**Лекция 8:**

**Метод многомерного моделирования**

**Цель лекции**

Изучив материал настоящей лекции, вы будете знать:

* основные элементы *многомерной модели*: *атрибуты*, *измерения*, *факты*, *гранулированность* ;
* основные схемы данных *многомерной модели*;
* что такое *аддитивные*, *полуаддитивные* и *неаддитивные факты* ;
* что такое *медленно меняющиеся измерения*, *быстро меняющиеся измерения*, *вырожденные измерения* ;
* что такое *иерархия измерений* ;
* что такое *схема "звезда"* ;
* что такое *схема "снежинка"* ;

и научитесь:

* моделировать *транзакционные таблицы фактов*, *таблицы фактов периодических моментальных снимков*, *таблицы фактов кумулятивных моментальных снимков* ;
* моделировать таблицы *измерений* и разрабатывать *иерархия измерений* ;
* использовать основные приемы разрешения отношения "многие ко многим" в *многомерной модели*.

**Основные понятия метода многомерного моделирования**

*Многомерное моделирование* (*Dimensional* *modeling*) проще для понимания, чем ER-*моделирование*. ***Многомерное моделирование*** **является методом моделирования и визуализации данных как множества числовых или лингвистических показателей или параметров (measures), которые описывают общие аспекты деятельности организации**. Как правило, при *многомерном моделировании* основное внимание фокусируется на числовых данных, таких как число продаж, баланс, *прибыль*, *вес*, или на объектах, которые можно пересчитать, таких как статьи, патенты, книги.

*Многомерное моделирование* имеет много общего с моделированием методом "*сущность-связь*" для реляционной модели, но отличается целями. Реляционная модель акцентируется на целостности и эффективности ввода данных. *Многомерная модель* (*Dimensional* *model*) ориентирована в первую *очередь* на выполнение сложных запросов к *БД*.

Метод *многомерного моделирования* базируется на следующих основных понятиях: *факты*, *атрибуты*, *измерения*, **параметры** (метрики), **иерархия**, *гранулированность*.

***Факт*** (*fact*) **— это набор связанных элементов данных, содержащих метрики и описательные данные**. Каждый *факт* обычно представляет *элемент данных*, численно описывающий *деятельность* организации, бизнес-операцию или событие, которое может быть использовано для анализа деятельности организации или бизнес-процессов. В ХД *факты* сохраняются в базовых таблицах реляционной *БД*. Например, *стоимость* товара, количество единиц товара и т.д.

***Атрибут*** (Аttribute) **– это описание характеристики реального объекта предметной области**. Как правило, *атрибут* содержит заранее известное *значение*, характеризующее *факт*. Обычно *атрибуты* представляются текстовыми полями с дискретными значениями. Например, габариты упаковки товара, запах товара.

***Измерение*** (*dimension*) **— это интерпретация факта с некоторой точки зрения в реальном мире**. *Измерения*, подобно *атрибутам*, содержат текстовые значения, которые сильно связаны по смыслу между собой. Обычно *измерения* представляются как оси многомерного пространства, точками которого являются связанные с ними *факты*. В *многомерной модели* каждый *факт* связан с одной или несколькими осями. *Измерения* обычно представляют нечисловые, *лингвистические переменные*, такие как филиалы организации, сотрудники организации, покупатели и т.д.

Например, при анализе продаж продукции, производимой или продаваемой организацией, такими *измерениями* обычно вступают *время*, покупатели, продавцы, *место* продажи или складирования товара.

*Измерения* задаются перечислением своих **элементов** (members). **Элемент измерения (dimensional member) — уникальное имя или идентификатор (лингвистическая переменная), используемая для определения позиции элемента**. Например, *измерение* " *Время* " может содержать следующие элементы: "все месяцы", "кварталы", "годы".

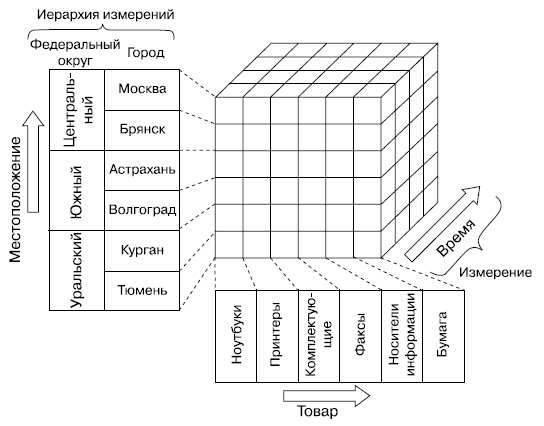
Часто элементы *измерения* находятся **в отношении "часть-целое" или "родитель-потомок"**, что позволяет ввести на *измерении* одну или несколько *иерархий*. Каждая *иерархия* может иметь несколько **уровней иерархии** (*hierarchy* levels). Каждый элемент *измерения* должен принадлежать только одному уровню иерархии, порождая таким образом *разбиение* на непересекающиеся подмножества. Примером может служить *иерархия* на *измерении* " *Время* ": год, полугодия, кварталы, месяцы и дни. Элемент *измерения* "неделя" может принадлежать двум месяцам, поэтому для него следует определить другую иерархию.

**Параметр, метрика или показатель (measure) — это числовая характеристика факта, который определяет эффективность деятельности или бизнес-действия организации с точки зрения измерения**. Как правило, *метрика* содержит заранее не известное *значение* характеристики *факта*. Конкретные значения метрики описываются с помощью переменных. Например, пусть метрикой является *численное выражение* продаж товара в деньгах, количество проданных единиц товара и т.д. *Метрика* определяется с помощью комбинации элементов *измерения* и, таким образом, представляет *факт*.

***Гранулированность*** (*Granularity*) **– это уровень детализации данных, сохраняемых в ХД**. Например, ежедневные объемы продаж.

**Многомерная модель**

*Многомерная модель* визуально представляется с помощью куба (или в случае более трех *измерений* — *гиперкуба*). Рассмотрим пример. Пусть объем продаж торговой организации есть *функция* от переменных "Товары", "Месяц" и "Регион продаж". Тогда в качестве *измерений* будут выступать "Товары", " *Время* " и "Месторасположение". На [рис. 9.1](https://intuit.ru/studies/courses/599/455/lecture/10167?page=1#image.9.1) приведен многомерный *куб данных* для представления данной функции.



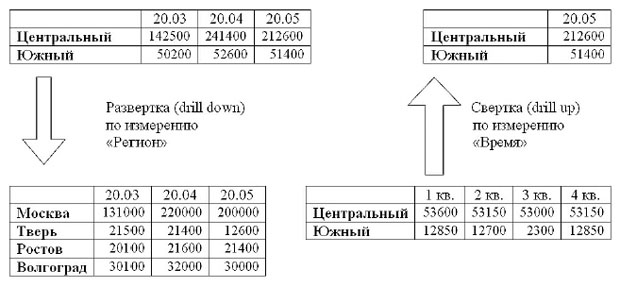
**Рис. 9.1.**Куб данных

На этих *измерениях* могут быть заданы следующие иерархии:

* *измерение* "Товары" — "Производитель-Категория-Товар";
* *измерение* "Месторасположение" — "Федеральный округ-Город-Магазин";
* *измерение* " *Время* " — "Год-Квартал-Месяц" или "Неделя-День".

*Многомерное моделирование* является основным методом логического проектирования ХД для OLAP-приложений. Для таких приложений типично выполнение операций свертывания и развертывания данных.

**Развертка** (*drill* down) и **свертка** (*drill* up) являются операциями перемещения вниз и вверх по уровням *иерархии измерения*. При выполнении развертки *пользователь* перемещается от верхних уровней к нижним уровням, которые содержат обычно более подробные данные. При выполнении свертки *пользователь* перемещается от нижних уровней иерархии к верхним, тем самым обобщая информацию на каждом уровне. При выполнении этих операций *путь* навигации определяется *иерархиями измерений*.

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_02_1.jpg)

**Рис. 9.2.**Операции свертки и развертки на кубе данных

Таким образом, иерархии задают пути суммирования объема продаж. Например, какой объем продаж товара HP LaserJet 1010 категории "Принтеры" производителя HP был на 3 неделе ноября в городе Брянске в магазине "Вычислительная техника".

Чтобы *приложение* могло использовать перечисленные *операции* для обработки данных, необходимо размещать данные в ХД определенным образом, т.е. поддерживать *многомерную модель*

**Факты**

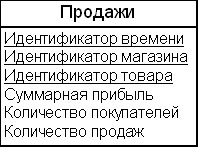
С точки зрения взаимосвязи *измерений* и *фактов* последние можно разбить на следующие классы:

* ***аддитивные факты*** (Additive facts). **Факт называется аддитивным, если его имеет смысл использовать с любыми** *измерениями* **для выполнения операций суммирования с целью получения какого-либо значимого результата**. Например, дискретные числовые показатели активности деятельности, такие как количество продаж, объем продаж и т.д.;
* ***полуаддитивные факты*** (Semiadditive facts). **Факт называется полуаддитивным, если его имеет смысл использовать совместно с некоторыми** *измерениями* **для выполнения операций суммирования с целью получения какого-либо значимого результата**. Например, числовые показатели интенсивности, такие как остаток на счете, уровень запасов на складе и т.д.;
* ***неаддитивные факты*** (Non-additive facts). **Факт называется неаддитивным, если его не имеет смысла использовать совместно с каким-либо** *измерением* **для выполнения операций суммирования с целью получения какого-либо значимого результата**. Например, *измерение* комнатной температуры;
* **числовые меры интенсивности** (Numerical Measures of *Intensity*). **Факт называется числовой мерой интенсивности, если он, являясь** *неаддитивным* **по** *времени* **, допускает агрегацию и суммирование по некоторому числу временных периодов**. Например, остаток на счете.

Рассмотрим примеры *неаддитивных фактов*. *Факты*, представляющие текстовые *атрибуты*, являются неаддитивными. Суммировать такие *факты* смысла не имеет. Для таких *фактов* имеет смысл подсчет их количества в таблице *фактов*. Цена единицы товара также является *неаддитивным фактом*. Не имеет смысла суммировать цену электрической розетки или себестоимость единицы продукции. Однако если сохраняется *произведение* стоимости единицы товара на количество проданного товара, то эта величина будет уже *аддитивным фактом*.

Проценты и отношения величин (валовая *прибыль*) являются *неаддитивными фактами*. Можно хранить как отдельные параметры *числитель* и знаменатель отношения, когда их раздельное суммирование имеет смысл. И это будут уже *аддитивные факты*. К *неаддитивным фактам* относятся также статистические средние суммы, такие как, например, средняя температура за день. Сумма средних дневных температур за неделю не имеет никакого смысла.

Проиллюстрируем понятие *аддитивного факта* на примере. Пусть сущность таблицы *фактов* "Продажи" ([рис. 9.3](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#image.9.3)) с *атрибутами* первичного *ключа* "*Идентификатор* времени", "*Идентификатор* товара", "*Идентификатор* магазина" имеет следующие *факты* (метрики): "Количество покупателей", "Суммарная *прибыль*" и "Количество продаж". Поскольку суммирование может быть выполнено для числовых метрик сущности "Количество продаж" и "Суммарная *прибыль*" по всем *измерениям*, эти *факты* являются *аддитивными*, как показано в примере ниже.



**Рис. 9.3.**Пример сущности "Продажи" с аддитивными и полуаддитивными фактами

**Пример 9.1**. *Аддитивные факты* можно суммировать по всем *измерениям*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| По измерению "Время": | | | | | | | | | |
| **Дата** | **Товар** | **Магазин** | | **Количество продаж** | | **Количество покупателей** | | **Суммарная прибыль** | |
| 23.01.2009 | CD диск | Компьютер | | 10 | | 10 | | 1500 | |
| 24.01.2009 | CD диск | Компьютер | | 35 | | 30 | | 5250 | |
| 25.01.2009 | CD диск | Компьютер | | 20 | | 15 | | 3000 | |
|  |  |  | | **65** | | **55** | | **9750** | |
| По измерению "Товар": | | | | | | | | | |
| **Дата** | **Товар** | | **Магазин** | | **Количество продаж** | | **Количество покупателей** | | **Суммарная прибыль** |
| 23.01.2009 | CD диск | | Компьютер | | 10 | | 6 | | 1500 |
| 23.01.2009 | Принтер | | Компьютер | | 1 | | 1 | | 5000 |
| 23.01.2009 | Сканер | | Компьютер | | 2 | | 2 | | 3000 |
|  |  | |  | | **13** | |  | | **9500** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| По измерению "Магазин": | | | | | |
| **Дата** | **Товар** | **Магазин** | **Количество продаж** | **Количество покупателей** | **Суммарная прибыль** |
| 23.01.2009 | CD диск | Компьютер | 10 | 10 | 1500 |
| 23.01.2009 | CD диск | Принтеры | 10 | 5 | 1500 |
| 23.01.2009 | CD диск | Оргтехника | 20 | 7 | 3000 |
|  |  |  | **40** | **22** | **6000** |

Проиллюстрируем понятие *полуаддитивного факта* на примере. Рассмотрим сущность *таблица* *фактов* "Продажи" ([рис. 9.3](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#image.9.3)). Как видно из примера 7.1, по *измерению* "Товары" суммирование по метрике "Количество покупателей" не выполнялось, а по *измерениям* " *Время* " и "Магазин" суммирование выполнялось. Этот *факт* является *полуаддитивным* по отношению к *измерению* "Товары".

**Пример 9.2**. *Полуаддитивные факты* не имеет смысла суммировать по некоторым *измерениям*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Суммирование метрики "Количество покупателей" по измерению "Товары": | | | | | |
| **Дата** | **Товар** | **Магазин** | **Количество продаж** | **Количество покупателей** | **Суммарная прибыль** |
| 23.01.2009 | Бумага для факсов | Компьютер | 10 | 6 | 1000 |
| 23.01.2009 | Бумага для принтера | Компьютер | 12 | 7 | 1320 |
|  |  |  |  | **13** |  |

В примере 9.2 показана *таблица*, в которой выполнено суммирование метрики "Количество покупателей" по *измерению* "Товары". Зададим следующий вопрос: "Сколько покупателей купили либо бумагу для принтера, либо бумагу для факса?". Ответ: где-то между 5-ю и 13-ю. Невозможность разделить количество покупателей между товарами для приведенной сущности делает этот *факт* *полуаддитивным* для *измерения* "Товары".

**Ключи в таблицах фактов**

Первичный *ключ* в таблице *фактов* является, как правило, составным первичным *ключом*. Он состоит из *множества* внешних *ключей*, которые служат первичными *ключами* *измерений*, связанных с *фактами*.

Следует заметить, что иногда несколько комбинаций внешних *ключей* представляют уникальный первичный *ключ* таблицы *фактов*. А с другой стороны, не всегда комбинация всех внешних *ключей* *измерений* гарантирует уникальность первичного *ключа* таблицы *фактов* (особенно если в таблице *фактов* есть *вырожденные измерения*, о которых см. ниже в этой лекции).

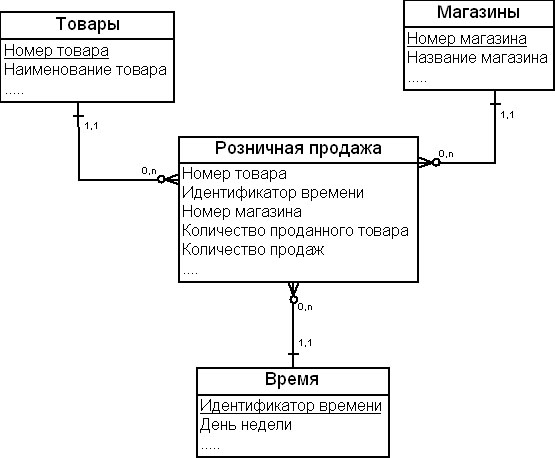
Рассмотрим подробнее вопрос уникальности первичного *ключа* таблицы *фактов*.

*Факты*, как правило, гранулированы. *Гранулированность* *фактов* определяет семантический смысл значения *факта* с точки зрения уровня детализации, связываемой с *фактом* информации. Например, общий объем продаж по данному магазину на указанный день по данному виду товара.

Часто *гранулированность* определяют на уровне отдельной бизнес-*операции*. Например, *продажа* товара. Иногда требуется сохранять информацию о продаже данного товара за день (здесь уже присутствует некоторый уровень агрегации данных).

Уникальность первичного *ключа* таблицы *фактов* гарантируется определением *гранулированности* *фактов*, а не условием уникальности внешних *ключей* от таблиц *измерений*, если *гранулированность* *фактов* не определена должным образом.

Рассмотрим таблицу *фактов* на схеме [рис. 9.4](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#image.9.4). *Гранулированность* *фактов* в ней составляет один день. *Факты* определяются для этой таблицы как суммы по всем операциям за день.

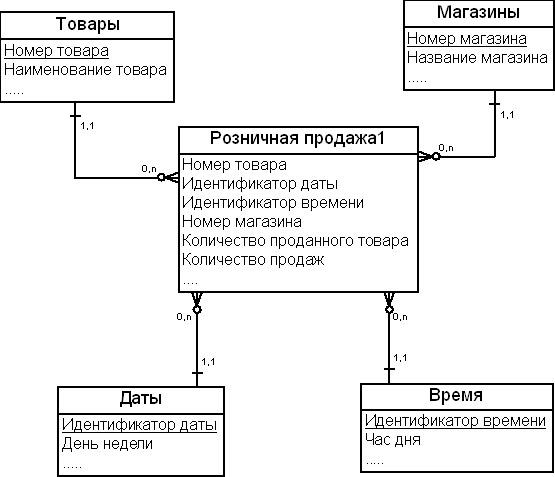


**Рис. 9.4.**Пример схемы данных для иллюстрации уникальности первичного ключа таблицы фактов: гранулированность фактов — одни сутки

Предположим, что в *измерении* "Товары" имеется два вида досок — С1 и С2. *Измерение* "Даты" содержит по одной строке за день. Предположим, что доски вида С1 были проданы 1 марта 2009 года в 9 утра в количестве 4 штук и вечером перед закрытием магазина в количестве 5 штук. Поскольку *гранулированность* *фактов* есть день, то в таблице *фактов* "Розничные продажи" для каждого проданного товара в каждом магазине будет вставлена только одна строка за день. Так, за 1 марта 2009 года в *поле* "Количество продаж" будет стоять число 9.

Если мы определим первичный *ключ* таблицы *фактов* как конкатенации первичных *ключей* *измерений* {Номер товара, *Идентификатор* даты, Номер магазина}, то этот первичный *ключ* будет уникальным, т.к. *фактом* будет только одна строка, вставленная для каждого товара, который продан в каждом магазине в определенный день. Уникальность первичного *ключа* гарантирована определением *гранулированности* *факта*.

Предположим, что руководству торговой компании необходимы отчеты почасовой продажи товаров. Тогда может быть использована схема, показанная на [рис. 9.5](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#image.9.5).



**Рис. 9.5.**Пример схемы данных для иллюстрации уникальности первичного ключа таблицы фактов: гранулированность фактов – один час

Предположим, что в течение 12 марта 2009 года доски вида С1 продавались следующим образом ([табл. 9.1](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#table.9.1)).

|  |  |
| --- | --- |
| Таблица 9.1. Данные по продажам досок | |
| **Время** | **Количество** |
| 8-00 | 5 |
| 8-30 | 5 |
| 11-00 | 2 |
| 16-00 | 3 |
| 18-00 | 2 |

Так как *гранулированность* таблицы *фактов* есть одна строка для каждого продукта, проданного в каждом магазине за каждый час, в таблицу будет вставлено 4 строки, фиксирующие продажу досок вида С1 за 2 марта 2009 года. Если мы определим первичный *ключ* таблицы *фактов* как конкатенации первичных *ключей* *измерений* {Номер товара, *Идентификатор* времени, *Идентификатор* даты, Номер магазина}, то этот первичный *ключ* будет уникальным.

Таким образом, *гранулированность* *факта* гарантированно определяет уникальность первичного составного *ключа* таблицы *фактов*.

**Таблицы фактов**

*Факты* в *многомерной модели* принято представлять в виде таблицы *фактов*. В логической модели "*сущность-связь*" *таблица* *фактов* представляется сущностью, *атрибутами* которой являются *факты* (метрики или описания) и составной *ключ*, связывающий таблицу *фактов* с таблицами *измерений* взаимосвязью "один ко многим".

Таблицы *фактов* разделяют на три основные категории, в зависимости от уровня детализации *фактов* ( *гранулированности* ).

* ***Транзакционная таблица фактов***. В такой таблице *фактов* сохраняют *факты*, которые фиксируют определенные события (транзакции). Это *факты*, описывающие каждое событие бизнеса. Например, продажи товара.
* ***Таблица фактов периодических моментальных снимков*** В такой таблице собирают *факты*, фиксирующие текущее состояние определенного направления бизнеса. Это *факты*, которые описывают текущее состояние определенного направления бизнеса для любой комбинации значений *измерений* за данный период *времени*. Например, продажи организации на определенную дату (ежедневно).
* ***Таблица фактов кумулятивных моментальных снимков*** В такой таблице собирают *факты*, фиксирующие некоторое итоговое состояние определенного направления бизнеса на текущий момент *времени*. Это *факты*, которые описывают промежуточные итоги деятельности организации по определенному направлению бизнеса для любой комбинации значений *измерений* за данный период *времени*. Например, продажи этого года на определенную дату.

Основными характеристиками таблицы *фактов* являются следующие.

1. Таблица *фактов* содержит числовые параметры (метрики).
2. Каждая таблица *фактов* имеет составной *ключ*, состоящий из первичных *ключей* таблиц *измерений*. Первичный *ключ* таблицы *измерений* является внешним *ключом* в таблице *фактов*.
3. Таблица *фактов* имеет, как правило, небольшое количество полей, не более 20-ти.
4. Данные в таблице *фактов* обладают следующими свойствами:
   * числовые параметры используются для агрегации и суммирования;
   * значения данных должны обладать свойствами аддитивности или полуаддитивности по отношению к *измерениям*, для того чтобы их можно было суммировать;
   * все данные таблицы *фактов* должны быть однозначно идентифицированы через *ключи* таблиц *измерений*, чтобы обеспечить доступ к ним через таблицы *измерений*.

Таким образом, таблицу *фактов* можно разделить на две части. Первая часть состоит из первичных *ключей* *измерений*, а вторая — из числовых параметров функционально зависящих от *ключей* таблиц *измерений*. Часто *таблица* *фактов* имеет суррогатный первичный *ключ*, сгенерированный системой, от которого по определению функционально полно зависят все параметры.

Пример таблицы *фактов* приведен на [рис. 9.3](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#image.9.3).

**Измерения**

Основными характеристиками таблицы *измерений* являются следующие.

1. Таблицы *измерений* содержат данные о детализации *фактов*.
2. Таблицы *измерений* содержат описательную информацию о числовых значениях в таблице *фактов*, т.е. они содержат *атрибуты* *фактов*.
3. Как правило, денормализованные таблицы *измерений* содержат большое количество полей.
4. Таблицы *измерений* содержат обычно значительно меньше строк, чем таблицы *фактов*.
5. *Атрибуты* таблиц *измерений* обычно используются при визуализации данных в отчетах и запросах.

На [рис. 9.6](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.6) приведен пример таблицы *измерения* " *Время* ". "*Идентификатор* времени" является первичным *ключом* таблицы *измерений*. Остальные поля являются *атрибутами* параметров таблицы *фактов*, зависящих от *времени*.



**Рис. 9.6.**Таблица измерений

Как правило, *таблица* *измерения* имеет первичный *ключ* и *атрибуты*, описывающие *факты* с точки зрения некоторого направления деятельности организации.

**Основные схемы многомерной модели**

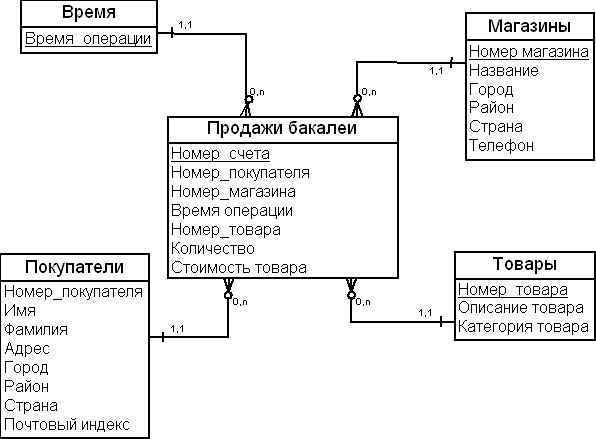
Существуют несколько схем для *многомерного моделирования* данных. Две из них считаются основными: *схема "звезда"* (*star schema*) и *схема "снежинка"* (*snowflake schema*). В более сложных случаях используются так называемые "многозвездочные" схемы или **схема с несколькими таблицами фактов**.

*Схема "звезда"* **имеет одну таблицу фактов и несколько таблиц измерений. Таблицы измерений являются денормализованными**.

*Схема "снежинка"* **имеет одну таблицу фактов и несколько нормализованных таблиц измерений**.

*Схема "звезда"* имеет те же самые элементы, что диаграммы "*сущность-связь*". Это — сущности, *атрибуты*, первичные и внешние *ключи*, взаимосвязи, кардинальность связи.

На [рис. 9.7](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.7) приведен пример *схемы "звезда"*, созданной для учета продажи бакалейных товаров. *Таблица* *фактов* "Продажи бакалеи" (учет операций продажи бакалейных товаров торговой компании) имеет один первичный *ключ* "Номер счета", четыре внешних *ключа* (по числу *измерений* ) и два параметра: "Количество" (проданного бакалейного товара) и "*Стоимость* товара".



**Рис. 9.7.**Схема "звезда"

*Измерение* " *Время* " является одним из критических элементов модели ХД. Если данные в OLTP-системах запросы фокусируются на текущем моменте *времени*, то в системах поддержки *принятия решений* (DSS-системах), для которых проектируются и создаются ХД, запросы фокусируются на задачах анализа данных, а именно — как данные изменялись в различные периоды *времени*. Например, каков был объем продаж торговой компании за последний квартал, месяц или каковы тенденции в покупках товаров в течение последнего квартала.

*Измерение* "Магазин" позволяет сгруппировать *операции* продаж по магазинам с учетом их географического положения.

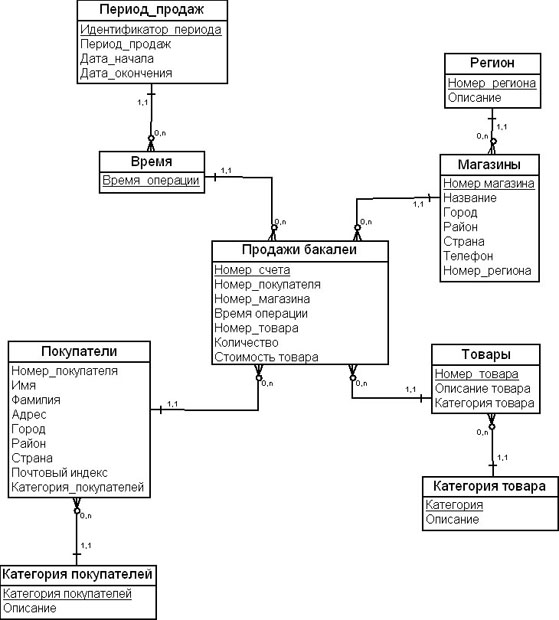
*Измерение* "Товары" позволяет анализировать типовые схемы закупок товаров и отвечать на вопрос, какие товары, как правило, покупаются одновременно покупателями.

*Измерение* "Покупатели" позволяет анализировать покупки с учетом их частоты, географического положения и количества.

Таблицы *измерений* являются своеобразным путеводителем при выборке строк из таблицы *фактов*.

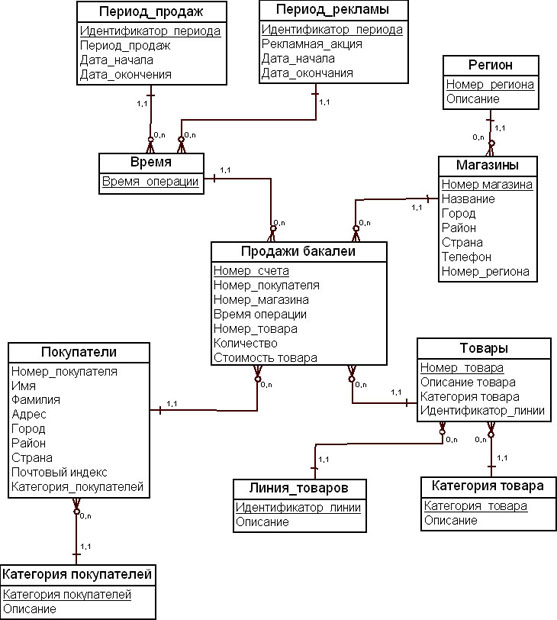
Таблицы *фактов* имеют большое количество строк. Таблицы *измерений*, как правило, имеют гораздо меньшее число строк. Одним из главных преимуществ использования такой схемы является то, что *производительность* *операции* *СУБД* соединения таблицы *фактов* и таких таблиц соединений будет близка к оптимальной.

*Схема "снежинка"* ([рис. 9.8](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.8)) добавляет иерархию в таблицы *измерений*. Например, *измерение* "Регион" группирует магазины по географическим регионам, *измерение* "Категория товара" группирует товары по категориям, *измерение* "Категория покупателей" группирует покупателей по категориям, а *измерение* "Период продаж" группирует продажи по периодам *времени*.

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_08.jpg)

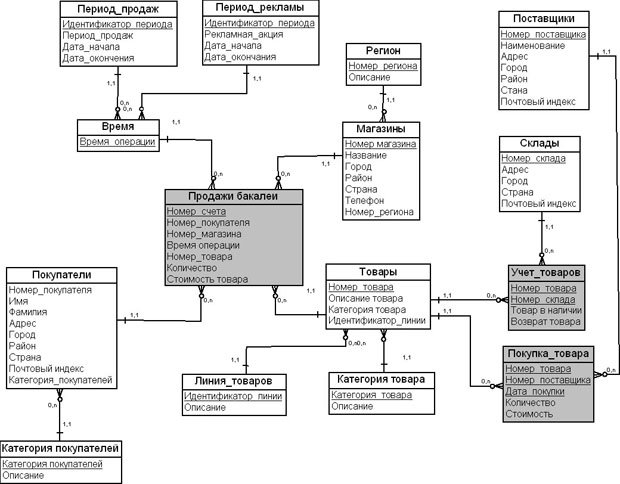
**Рис. 9.8.**Схема "снежинка"

Таким образом, использование иерархий превращает *схему "звезда"* в *схему "снежинка"*. Вы можете расширять *схемы "снежинка"* за счет добавления в нее новых иерархий, как показано на [рис. 9.9](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.9).

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_09.jpg)

**Рис. 9.9.**Добавление в схему "снежинка" новых иерархий

На [рис. 9.10](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.10) приведен пример схемы с несколькими таблицами *фактов* – схемы "созвездие *фактов* " (*Fact* Constellation *Schema*). На схеме имеет три таблицы *фактов*: "Продажи бакалеи", "Учет\_товара" и "Покупка\_товара", у которых имеются как общие *измерения* ("Товары"), так и эксклюзивные (например, *измерение* "Склады" для таблицы *фактов* "Учет\_товара").

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_10.jpg)

**Рис. 9.10.**Схема "созвездие фактов"

### ТепМоделирование таблиц фактов

При моделировании таблиц *фактов* проектировщик ХД в первую *очередь* опирается на рассмотрение такой характеристики, как степень детализации ( *гранулированности* ) *фактов*.

Выше мы рассмотрели таблицы *фактов*, которые содержат параметры (метрики) на самом низком уровне детализации ( *гранулированности* ), таком как уровень бизнес-*операции* организации, или мгновенной транзакции. Мгновенная *транзакция* предоставляет возможность зарегистрировать описание *факта* в определенный момент *времени*. Это есть ***транзакционные таблицы фактов***. Пример такой таблицы приведен на [рис. 9.3](https://intuit.ru/studies/courses/599/455/lecture/10167?page=2#image.9.3). *Факты* "Суммарная *прибыль*", "Количество продаж" и "Количество покупателей" фиксируются в таблице *фактов* на уровне *операции* продажи товара (мгновенной транзакции).

Обратим внимание на то обстоятельство, что такие таблицы *фактов* могут содержать тестовые *атрибуты* и неаддитивные числовые *факты*.

С каждым уровнем *гранулированности* может быть связан агрегат. **Агрегатами являются суммы значений параметров или статистические функции от значений параметров, взятые на определенном уровне детализации** ( *гранулированности* ). Агрегаты на одних и тех же *фактах* можно определять на одном или более уровнях детализации.

Обычно в ХД используют два типа таблиц агрегатов *фактов*, как было указано в этой лекции выше:

1. со степенью детализации на уровне периодического снимка данных, представляющего промежуток *времени* заданной продолжительности ( *таблица фактов периодических моментальных снимков* );
2. со степенью детализации на уровне аккумулирующего снимка, представляющего всю историю *фактов* (исторические данные) с заданного и до текущего моментов *времени* ( *таблица фактов кумулятивных моментальных снимков* ).

Агрегаты имеют большое *значение* для производительности запросов. *Запрос* будет выполнен гораздо быстрее на 10-ти заранее вычисленных строках таблицы *фактов*, чем на десяти тысячах строк более низкого уровня детализации.

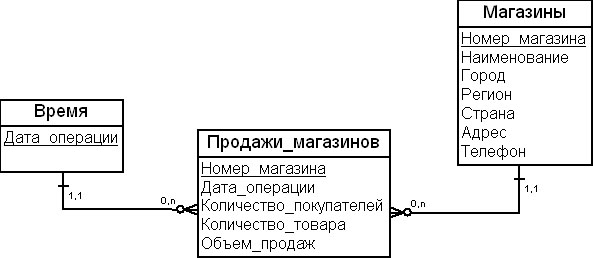
Например, для схемы на [рис. 9.6](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.6) можно построить агрегат "ежедневные продажи" по магазинам или регионам, можно построить агрегат "еженедельные продажи" по категориям товара.

**Таблицей агрегатов фактов (Aggregate fact table) называется таблица фактов, которая содержит агрегаты некоторых фактов модели**.

Проектировщик добавляет таблицы агрегатов *фактов* в *многомерную модель* данных и устанавливает ее связи с *измерениями*.

Построим таблицу агрегатов *фактов* "Продажи магазинов", содержащую агрегаты *фактов* по ежедневным продажам магазина, как показано на [рис. 9.11](https://intuit.ru/studies/courses/599/455/lecture/10167?page=4#image.9.11). Эта *таблица* содержит итоговые суммы по ежедневным продажам каждого магазина.

* Поле "Количество покупателей" указывает, как много покупателей сделали покупки.
* Поле "Количество товаров" указывает, сколько единиц товара было продано.
* Поле "Объем продаж" указывает количество товара, проданного за день.

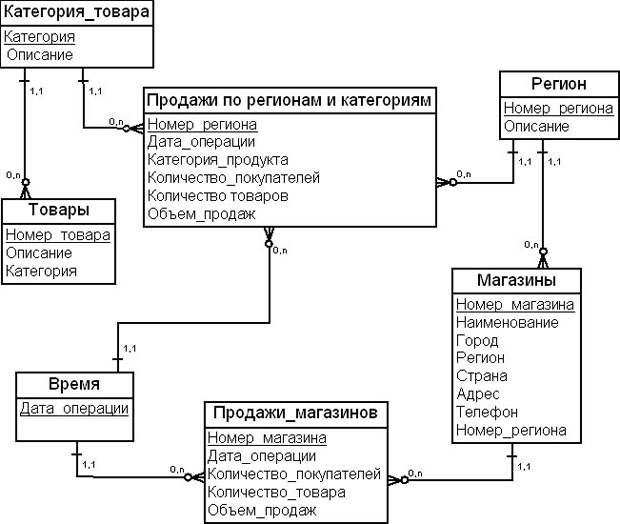


**Рис. 9.11.**Таблица агрегатов фактов "Продажи магазинов"

Это пример *таблицы фактов периодических моментальных снимков*.

На основе этой таблицы агрегатов *фактов* могут быть построены *производные* агрегаты, такие как среднее количество покупок, сделанное одним покупателем, или средняя сумма, затраченная одним покупателем на покупки товаров.

Проектировщик может спроектировать другую таблицу агрегатов *фактов*, которая содержит данные о ежедневных продажах по регионам и категориям продуктов, как показано на [рис. 9.12](https://intuit.ru/studies/courses/599/455/lecture/10167?page=4#image.9.12).

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_12.jpg)

**Рис. 9.12.**Таблица агрегатов фактов "Продажи магазинов"

Можно добавлять *произвольное* количество таблиц агрегатов *фактов* в *многомерную модель* ХД. Заметим, что таблицы агрегатов *фактов* представляют собой заранее заготовленные ответы на известные вопросы. Незапланированные запросы к агрегатам, как правило, не поддерживаются.

Для таблицы *фактов*, связанной с 10-ю таблицами *измерений*, потенциально существует более трех миллионов таблиц агрегатов *фактов*. Какие таблицы агрегатов *фактов* должны быть построены в модели данных, зависит от поставленной *цели проекта* ХД и проектных решений проектировщика ХД.

При определении агрегатов полезно пользоваться принципом Парето: 20% возможных кандидатов в агрегаты будут действительно востребованы пользователями.

*Таблица фактов кумулятивных моментальных снимков* агрегирует в строке таблицы *фактов* данные за все *время* существования какого-либо события. Например, *таблица* *фактов* кумулятивных снимков может хранить информацию по зарплатным картам сотрудников какой-либо организации. Как правило, такие карты выдаются банком на срок от двух до пяти лет, а потом меняются на новые карты.

В качестве такой таблицы агрегатов рассмотрим компанию, которая занимается трудоустройством. Предположим, что ежедневно компания рекламирует в прессе более 10000 вакансий. Руководство компании хотело бы оценить эффективность своих сотрудников, которые нанимают людей на работу, а именно: сколько *времени* длится процесс трудоустройства кандидата, сколько *времени* тратится на ожидание поступлений *резюме*, сколько *времени* тратит *менеджер* компании на трудоустройство кандидата, который обратился за услугами в компанию.

В этой таблице *фактов* одна строка (факт) будет содержать три периода *времени* для каждого менеджера и для каждой вакансии. Пример схемы приведен на [рис. 9.13](https://intuit.ru/studies/courses/599/455/lecture/10167?page=4#image.9.13).

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_13.jpg)

**Рис. 9.13.**Таблица агрегатов фактов "Трудоустройство"

Такой вид таблиц агрегатов *фактов* используется, как правило, для кратковременных процессов и для процессов, основанных на случайных событиях, типа банковской транзакции.

В [табл. 9.2](https://intuit.ru/studies/courses/599/455/lecture/10167?page=4#table.9.2) приведено сопоставление основных видов таблиц *фактов*.

|  |  |  |  |
| --- | --- | --- | --- |
| Таблица 9.2. Сравнение видов таблиц фактов | | | |
|  | ***Транзакционная таблица фактов*** | ***Таблица фактов периодических моментальных снимков*** | ***Таблица фактов кумулятивных моментальных снимков*** |
| Определение *гранулированности* таблицы фактов | Одна строка на бизнес-операцию | Одна строка на период | Одна строка для всего периода завершенного события |
| Измерения | Используют факты на самом низком уровне детализации по измерению "дата/время" | Используют факты на самом некотором уровне агрегации по измерению "дата/время" (по концу периода) | Используют факты по нескольким измерениям "дата/время" для фиксации результатов в различных контрольных точках |
| Общее количество задействованных измерений | Больше, чем в таблицах фактов периодических снимков | Меньше, чем в *транзакционных таблицах фактов* | Наибольшее количество измерений для таблиц фактов |
| Факты | Факты связаны с операционной деятельностью | Факты связаны с периодической деятельностью | Факты связаны с деятельностью, которая имеет определенное время существования |
| Обновления | Не допускаются | Не допускаются | Допускаются |
| Кардинальность таблицы фактов | Растет быстро | Растет медленнее, чем в *транзакционных таблицах фактов* | Растет медленнее, чем в *таблицах фактов периодических моментальных снимков* |

Теперь обратимся к изучению вопроса моделирования *измерений*.

### Моделирование таблиц измерений

#### Медленно меняющиеся измерения

Значение полей таблиц *измерений* может изменяться со *временем*. Такие изменения приводят к проблемам в проектировании таблиц *измерений*. Это связано с тем, что ХД, по определению и своему назначению, предназначены хранить исторические данные и обеспечивать доступ пользователей к ним.

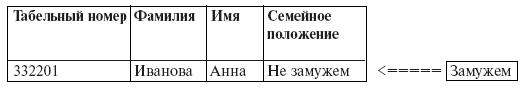
***Медленно меняющимися измерениями*** (Slowly Changing Dimensions) **называются таблицы измерений, в которых некоторые** *атрибуты* **могут изменить свои значения по истечении некоторого периода** *времени*, **причем частота таких изменений является небольшой**.

В качестве примера рассмотрим таблицы *измерений* "Покупатель" или "Товары". Предположим, что первичные *ключи* таблиц не меняются, а *атрибуты* таблицы *измерений* могут изменяться со *временем*. В таком случае выделяют три типа действий.

* Тип 1. Изменить значение *атрибута* таблицы *измерений* на новое значение. При этом будет потеряна хронология.
* Тип 2. Создать новую строку в таблице *измерений* с новым значением суррогатного *ключа*.
* Тип 3. Создать дополнительный *атрибут* таблицы *измерений* с новым значением.

В первом случае старое значение *атрибута* меняется на новое значение, как показано в примере ниже. Этот прием является простым в реализации, но не позволяет отследить историю изменений семейного статуса.

**Пример 9.3**. *Медленно меняющиеся измерения*. Тип 1



Во втором случае создается новая запись в таблице *измерения* с новым суррогатным *ключом* (пример 9.4). С точки зрения бизнес-операций организации у женщины, изменившей семейное положение, меняется табельный номер. При этом строки таблицы *фактов* "до замужества" будут относиться к строке таблицы *измерения* с табельным номером 332201, а "после замужества" – к строке таблицы *измерения* с табельным номером 332209.

**Пример 9.4**. *Медленно меняющиеся измерения*. Тип 2

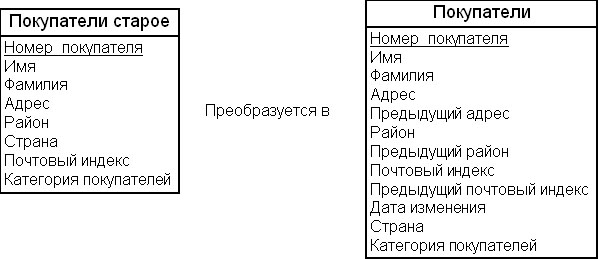
|  |  |  |  |
| --- | --- | --- | --- |
| **Табельный номер** | **Фамилия** | **Имя** | **Семейное положение** |
| 332201 | Иванова | Анна | Не замужем |
| 332209 | Иванова | Анна | Замужем |

В третьем случае создаются новые поля в таблице *измерения*, отражающие смену семейного положения, как показано в примере 7.5 ниже. Как видно из примера, потребовалось два дополнительных поля для сохранения предыдущего статуса семейного положения и даты его изменения.

**Пример 9.5**. *Медленно меняющиеся измерения*. Тип 3

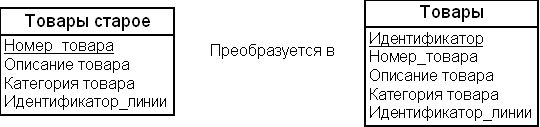
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Табельный номер** | **Фамилия** | **Имя** | **Предыдущее семейное положение** | **Текущее семейное положение** | **Дата изменений** |
| 332201 | Иванова | Анна | Не замужем | Замужем | 02.05.2009 |

Обратимся к схеме на [рис. 9.8](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.8). Рассмотрим *измерение* "Покупатели". Покупатель может изменить район проживания, и следовательно, изменятся *атрибуты* "Город", "Адрес" и "Почтовый индекс". Допустим, что организация анализирует *факты* продаж по регионам. Тогда *измерение* "Покупатели" может быть отнесено к типу 3. В этом случае необходимо изменить модель *измерения* "Покупатели", как показано на [рис. 9.14](https://intuit.ru/studies/courses/599/455/lecture/10167?page=5#image.9.14).



**Рис. 9.14.**Пример медленно меняющегося измерения типа 3

*Измерение* "Товары" на схеме [рис. 9.8](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.8) может быть отнесено к *медленно меняющимся измерениям* типа 2. Организация может перевести товар в другую категорию (при создании, например, новой категории товаров) и при этом хранить историю изменений, т.е. требуется изменить значение *атрибута* описание товара. В этом случае, как описано выше, создается новая запись. Обратим внимание на то, что если при этом номер товара не меняется, то нужно ввести в *измерение* суррогатный *ключ*, как показано на [рис. 9.15](https://intuit.ru/studies/courses/599/455/lecture/10167?page=5#image.9.15).



**Рис. 9.15.**Пример медленно меняющегося измерения типа 2

*Измерение* "Категория товара" на схеме [рис. 9.8](https://intuit.ru/studies/courses/599/455/lecture/10167?page=3#image.9.8) может быть отнесено к *медленно меняющимся измерениям* типа 1. Маркетологи торговой организации могут уточнить описание категории товара и новое описание должно фигурировать во всех отчетах.

Перечисленные типы практически полностью характеризуют *медленно меняющиеся измерения*. Можно комбинировать указанные типы между собой. Комбинация типов 1 и 2 будет нужна, когда для одних *атрибутов* создается новая запись (тип 2), а для других обновляется существующее значение (тип 1). Комбинация типов 2 и 3 будет востребована, когда необходимо не только сохранять историю изменений за счет создания новой записи (тип 2), но и иметь возможность рассматривать некоторые *факты* с точки зрения исторических значений некоторой записи *измерения* (тип 3).

При выборе типа *медленно меняющихся измерений* проектировщику ХД следует придерживаться следующей схемы принятия решений:

ЕСЛИ требуется сохранять историю измерения, ТО следует выбрать

тип 2.

В ПРОТИВНОМ СЛУЧАЕ

ЕСЛИ необходимо сравнивать текущее значение атрибута с перво-

начальным или предыдущим, ТО следует выбрать тип 3

В ПРОТИВНОМ СЛУЧАЕ

следует выбрать тип 1.

Суррогатные *ключи* предпочтительнее естественных *ключей* (бизнес-ключей), особенно в случае использования типа 2.

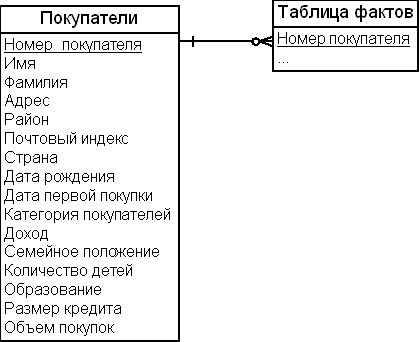
#### Быстро меняющиеся измерения

Значение *атрибутов* *измерений* может меняться часто за короткий период *времени*. Примером такого *измерения* может быть номенклатура изделий фабрики по выпуску проволоки. Заказчики часто требуют изготовление проволоки заданного диаметра. Как показывает практика, номенклатура изделий таких фабрик меняется приблизительно на треть в год.

***Быстро меняющимися измерениями*** (Rapidly Changing Dimensions) **называются таблицы измерений, в которых некоторые** *атрибуты* **могут часто менять свои значения в короткие периоды** *времени*.

Модели для управления такими *измерениями* зависят от кардинальности таблиц *измерений*. Если кардинальность таблиц *измерений* является небольшой (до 10000 записей), то может быть использован такой же подход, как в случае *медленно меняющихся измерений*. В случае очень больших таблиц *измерений* (до миллиона записей) следует избегать дублирования записей и не создавать новые дополнительные записи. Поскольку необходимо помнить о производительности выполнения запросов с такими *измерениями*, подходы к моделированию таких *измерений* являются важными.

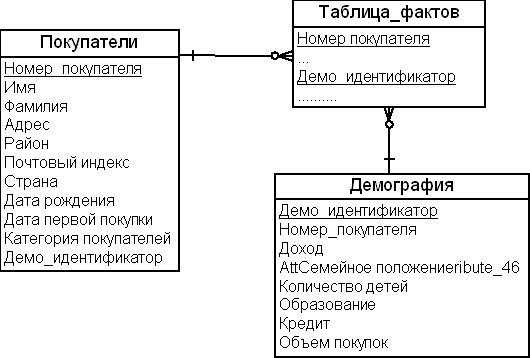
В качестве примера *быстро меняющихся измерений* рассмотрим *измерение* "Покупатели" ([рис. 9.16](https://intuit.ru/studies/courses/599/455/lecture/10167?page=5#image.9.16)). Для больших торговых компаний это *измерение* содержит значительное число записей и включает в себя большое количество *атрибутов*, в том числе демографических. Демографические *атрибуты* часто используются в исследовании предпочтений различных социальных групп, причем покупатели часто мигрируют между социальными группами.



**Рис. 9.16.**Пример быстро меняющегося измерения

Основным приемом моделирования *быстро меняющихся измерений* является логическое разбиение таблицы *измерения* на две или более таблицы. При этом для быстро меняющихся *атрибутов* часто используют дискретный диапазон изменений, чтобы сократить объем данных в таблице.

Суть приема логического разбиения состоит в следующем: создаются две сущности, одна из которых содержит *атрибуты*, которые меняются медленно, а другая сущность включает в себя *атрибуты*, которые меняются быстро. Для *измерения* "Покупатели", рассмотренного на [рис. 9.16](https://intuit.ru/studies/courses/599/455/lecture/10167?page=5#image.9.16), это может быть сделано, как показано на [рис. 9.17](https://intuit.ru/studies/courses/599/455/lecture/10167?page=5#image.9.17) ниже.



**Рис. 9.17.**Пример разбиения быстро меняющегося измерения

Далее можно предположить, что кредитные операции покупателя в торговой сети компании будут изменяться быстрее, чем остальные демографические данные. Тогда можно разбить *измерение* "Демография" на два, вынеся во второе *измерение* *атрибуты*, описывающие кредит и покупки в кредит.

Достоинством такого подхода является то, что частая смена профиля покупателя не приводит к сложности представления данных и не увеличивает размеров используемой памяти за счет отсутствия в схеме дублирования значений медленно меняющихся или неизменяемых *атрибутов*.

Использование *разбиения таблиц* имеет и свои недостатки, в частности, навигация по данным будет осуществляться медленнее, а применение диапазонов дискретных значений для демографических *атрибутов* непрактично.

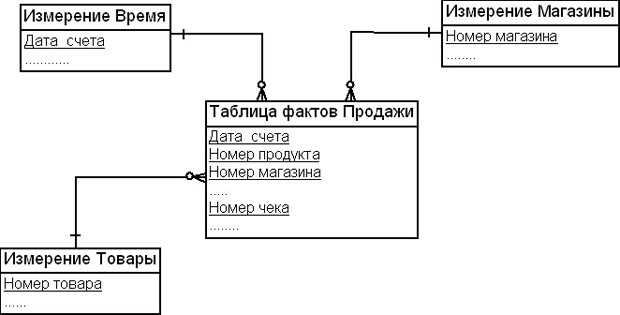
Заметим также, что если соединение между *измерениями* "Покупатели" и "Демография" будет иметь невысокую кардинальность, то придется либо разбивать таблицу *фактов*, чтобы обеспечить более эффективное соединение, либо вводить в *измерение* "Демография" внешний *ключ*, указывающий на *измерение* "Покупатели", как сделано в нашем примере выше.

#### Вырожденные измерения

***Вырожденным измерением*** (Degenerate Dimension) называется *ключ* в таблице *фактов*, по которому не производится соединение с таблицей, поскольку все связанные с этим *ключом* *атрибуты* размещаются в других *измерениях*. Это *измерение* без *атрибутов*. Обычно *вырожденное измерение* представлено *атрибутами* *ключа* *измерения* в таблице *фактов* без соответствующей таблицы *измерений*.

*Вырожденные измерения* часто встречаются в таблицах *фактов*, которые созданы на уровне детализации данных, соответствующем уровню мгновенной транзакции. Как правило, *вырожденными измерениями* являются идентификаторы, присваиваемые учетными системами, такие как номера заказов, номера транзакций по кредитным картам, номера счетов, номера счетов-фактур и т.п. *Вырожденные измерения* являются естественными *ключами* "родителей", соответствующих детальным записям.

Несмотря на то, что по *вырожденному измерению* не производится соединения с таблицей, содержащей дополнительные *атрибуты*, *вырожденное измерение* может быть полезно для группировки родственных записей в таблице *фактов*. Например, на [рис. 9.18](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.18) для розничной торговли *атрибут* "Номер чека" группирует все товары, купленные в одной корзине, и является *вырожденным измерением*.

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_18.jpg)

**Рис. 9.18.**Пример вырожденного измерения

В таблице *фактов* может быть несколько *вырожденных измерений*. Например, таблица *фактов* по *отгрузкам товаров* может содержать как номер заказа, так и номер товарной накладной.

*Вырожденные измерения* могут служить для поддержания связи с учетными системами, а также быть полезными на этапе разработки *ETL* для связи записей таблицы *фактов* с записями в учетных системах в целях тестирования или проверки целостности.

Значения *вырожденных измерений*, как правило, уникальны и имеют небольшие размеры. Однако если будет использован суррогатный *ключ*, скажем, для экономии памяти, или если идентификатор учетной системы выражен длинной строкой, то **измерение уже не является вырожденным**.

Один из практических критериев идентификации *вырожденных измерений* (на этапе эксплуатации) — наличие таблиц *измерений*, кардинальность которых растет пропорционально кардинальности таблиц *фактов*. Такие таблицы *измерений* следует сделать *вырожденными измерениями*.

#### Иерархии измерений

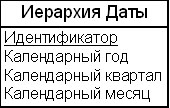
**Иерархией называется взаимосвязанный набор отношений "многие к одному", состоящий из последовательности уровней**. Например, иерархия "Регион" определяется уровнями "Область", "Район", "Город". Иначе говоря, иерархия является спецификацией уровней, которые представляют взаимосвязи между различными *атрибутами* в иерархии. Как правило, иерархии обогащают семантику взаимоотношений между данными в *многомерной модели*.

В *многомерном моделировании* различают три типа иерархий:

* сбалансированные иерархии (Balanced hierarchy);
* несбалансированные иерархии (*Unbalanced* hierarchy);
* иерархии с пропущенными уровнями (*Ragged* hierarchy).

**Сбалансированная иерархия – это иерархия, в которой все ветви** *измерения* **имеют одно и то же количество уровней**. Например, иерархия "Год" — "Квартал" —"Месяц". Каждый уровень имеет один и тот же тип и логически эквивалентен.

Сбалансированная иерархия состоит из фиксированного числа уровней. Число *атрибутов* в таблице *измерения* соответствует числу уровней в иерархии. На [рис. 9.19](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.19) приведен пример сбалансированной иерархии "Даты".

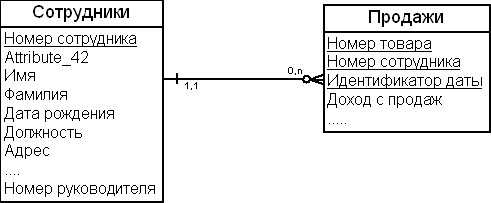


**Рис. 9.19.**Пример сбалансированной иерархии

Таблица *измерения* может иметь несколько *иерархий*. Например, в таблицу *измерения* на [рис. 9.19](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.19) можно добавить *атрибуты* иерархии "Финансовый год" "Финансовый квартал" — "Финансовый месяц".

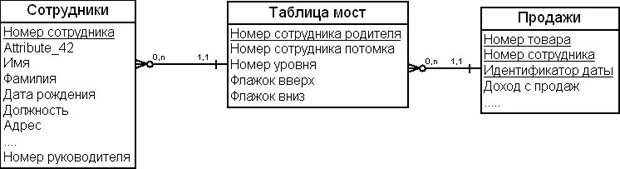
**Несбалансированная иерархия – это иерархия, в которой все ветви** *измерения* **имеют различное число уровней**. *Измерение* типа "родитель-потомок" является примером несбалансированной иерархии. Примером несбалансированной иерархии может быть организационная структура организации: "Директор" — "Заместители директора" — "Отделы" — "Лаборатории" – "Группы".

Общим подходом при моделировании взаимосвязей "родитель-потомок" в реляционных БД является введение *ключа* сущности потомка в сущность родителя. Этот *ключ* называется **рекурсивным указателем** (recursive pointer). Пример моделирования приведен на [рис. 9.20](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.20). *Атрибут* "Номер руководителя" является рекурсивным указателем, который указывает на "Номер сотрудника".



**Рис. 9.20.**Пример несбалансированной иерархии

Структура, приведенная на [рис. 9.20](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.20), не работает для некоторых запросов. Предположим, что нам нужно показать суммарный доход с продаж сотрудника с номером 2. Если сотруднику с номером 2 подчинены сотрудники с номерами 5, 4, 7, то нам нужно показать суммарный доход с продаж, полученный от сотрудников с номерами 2, 5, 4, 7. Как правило, при использовании предложения GROUP BY одного оператора SELECT невозможно провести суммирование вниз по уровням иерархии. Для разрешения этой ситуации используется вспомогательная *таблица-мост*, как показано на [рис. 9.21](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.21).

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_21.jpg)

**Рис. 9.21.**Пример использования таблицы-моста в отношении "родитель-потомок"

***Таблицей-мостом*** **называется вспомогательная таблица, которая предназначена упростить работу с рекурсивными отношениями, отношениями "многие ко многим", отношениями типа иерархии при использовании реляционной модели данных**.

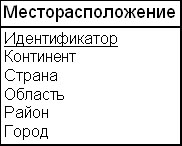
Рассмотрим, как работает *таблица-мост*. Таблица *измерений* "Сотрудники" имеет по одной строке на каждого сотрудника на любом уровне иерархии. Она содержит по одной строке для каждого пути в дереве иерархии — т.е. *таблица-мост* содержит одну строку для каждого пути иерархии от служащего к подчиненному ему служащему, а также строку, содержащую ссылку от служащего на самого себя. Кроме этого, *таблица-мост* содержит номер уровня иерархии (количество уровней между родителем и потомком в отношении), флажок "вниз", который указывает, что данный служащий находится на самом нижнем уровне иерархии, и флажок "вверх", который указывает, что он на самом верхнем уровне иерархии.

*Таблица-мост* может участвовать в операции соединения таблицы *измерения* "Сотрудники" и таблицы *фактов* "Продажи" двумя способами. При движении по иерархии вниз таблица *фактов* посредством *ключа* "Номер сотрудника" соединяется с *таблицей-мостом* по *ключу* "Номер сотрудника потомка", а таблица *измерения* "Сотрудники" посредством *ключа* "Номер сотрудника" соединяется с *таблицей-мостом* "Номер сотрудника родителя". Поле "Номер уровня" определяет глубину движения по иерархии. Поле "Флажок вниз" говорит, что достигнут самый низкий уровень несбалансированной иерархии. При движении вверх по иерархии таблица *фактов* посредством *ключа* "Номер сотрудника" соединяется с *таблицей-мостом* по *ключу* "Номер сотрудника родителя", а таблица *измерения* "Сотрудники" посредством *ключа* "Номер сотрудника" соединяется с *таблицей-мостом* "Номер сотрудника потомка". Поле "Номер уровня" также определяет глубину движения по иерархии. Поле "Флажок вверх" говорит, что достигнут самый высокий уровень несбалансированной иерархии.

Заметим, что использование *таблиц-мостов* усложняет как проектирование модели данных, так и ее последующее сопровождение. Для сопровождения таких таблиц, как правило, приходится разрабатывать пакеты из хранимых процедур.

**Иерархией с пропущенными уровнями (Ragged hierarchy) называется такая иерархия, в которой допускается отсутствие одного из уровней при заполнении ее данными**, т.е. различные варианты иерархии сохраняются в одной структуре данных.

Рассмотрим таблицу *измерений* "Месторасположение" на [рис. 9.22](https://intuit.ru/studies/courses/599/455/lecture/10167?page=6#image.9.22), которая представляет географическое расположение населенных пунктов.



**Рис. 9.22.**Пример структуры данных для реализации иерархии с пропущенными уровнями

Например, для населенных пунктов в Российской федерации заполнение таблицы *измерений* может быть следующим.

* Для большинства населенных пунктов Московской области будет иметь вид:

|  |
| --- |
| "Континент" = Евразия |
| "Страна" = Российская федерация |
| "Область" = Московская область |
| "Район" = Ногинский район |
| "Город" = Купавна |

* Для городов областного подчинения поле "Район" будет незаполненным:

|  |
| --- |
| "Континент" = Евразия |
| "Страна" = Российская федерация |
| "Область" = Московская область |
| "Район" = NULL |
| "Город" = Черноголовка |

#### Отношения "многие ко многим" в измерениях

Таблицы *измерений* могут находиться в отношении "многие ко многим" между собой. Например, поставщики могут поставлять товары на разные склады, а магазины получать товары с различных складов. Отношение "многие ко многим" может существовать между: 1) таблицей *измерения* и таблицей *фактов*, 2) между таблицами *измерений*.

Как правило, отношения между таблицами БД "многие ко многим" разрешаются путем создания дополнительной таблицы, которая управляет таким отношением (требование четвертой нормальной формы при проектировании реляционных БД).

В *многомерном моделировании* ХД для разрешения отношения "многие ко многим" между таблицами *измерений* могут быть использованы два типа таких дополнительных таблиц: **"пустая" таблица** *фактов* или *таблица фактов без метрик* (factless fact table) и *таблица-мост* (bridge table).

***Таблицей фактов без метрик*** **называется таблица фактов, которая не содержит числовых параметров или метрик**. Обязательным *атрибутом* этой таблицы является составной *ключ*, который состоит из первичных *ключей* сущностей, находящихся в отношении "многие ко многим". Такие таблицы *фактов* используются, как правило, для представления событий или охвата событий и содержат информацию о том, что что-то произошло или не произошло.

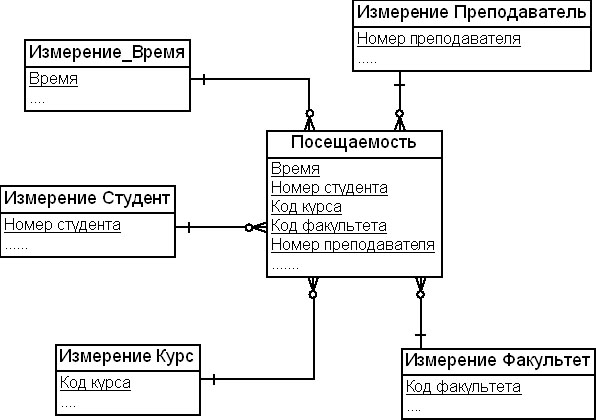
Различают два типа *таблиц фактов без метрик*: **таблицы фактов отслеживания событий** (event tracking tables) и **таблицы фактов охвата событий** (coverage tables).

Таблица *фактов* охвата событий содержит описание того, что еще не произошло, — например, какие товары еще не продавались по рекламной акции.

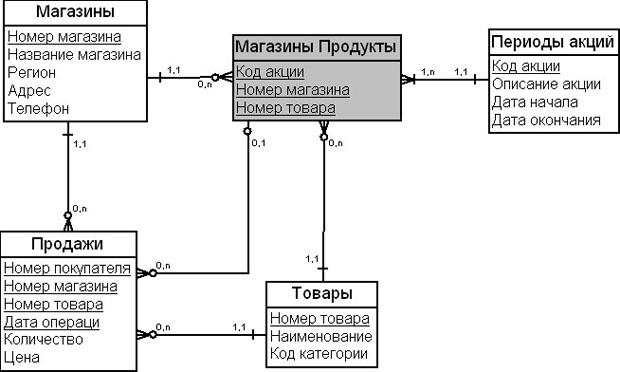
Таблица *фактов* отслеживания событий фиксирует событие, т.е. дату или *время* события и его описание.

В качестве примера рассмотрим посещаемость студентами курсов, читаемых различными преподавателями. Студенты изучают в университете различные учебные курсы, учебные курсы проводятся различными преподавателями, студенты учатся на разных факультетах и т.д. Сущности "Студент", "Преподаватель". "Факультет", "Курс" и " *Время* " находятся в отношении "многие ко многим". Мы можем построить *таблицу фактов без метрик* "Посещаемость" для отслеживания присутствия студентов различных факультетов на курсах различных преподавателей, как на [рис. 9.23](https://intuit.ru/studies/courses/599/455/lecture/10167?page=7#image.9.23). Таким образом, будут разрешены отношения "многие ко многим" у рассматриваемых выше сущностей ( *измерений* ).

Таблицу *фактов* охвата событий, или, как ее еще называют, таблицу существования (Existence Table), обычно используют, когда для таблиц *измерения* типа *времени* может существовать перекрывание периодов — например, в маркетинговой акции по продвижению товара на рынок, связанной с гибкой сеткой цен на товары. При этом не все магазины будут участвовать в продвижении товара в одно и то же *время*. Также не все товары будут участвовать в данной маркетинговой акции. Пример использования таблицы *фактов* охвата событий "Магазины Продукты" приведен на [рис. 9.24](https://intuit.ru/studies/courses/599/455/lecture/10167?page=7#image.9.24).



**Рис. 9.23.**Пример таблицы фактов без метрик

[](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_24.jpg)

[увеличить изображение](https://intuit.ru/EDI/20_07_20_2/1595197216-9970/tutorial/632/objects/9/files/07_24.jpg)  
**Рис. 9.24.**Пример таблицы фактов охвата событий

В приведенном на [рис. 9.16](https://intuit.ru/studies/courses/599/455/lecture/10167?page=5#image.9.16)а примере таблица *фактов* охвата событий используется для определения того, какой товар был продан в каком магазине в течение периода маркетинговой акции.

Такие таблицы должны помочь аналитикам в оценке эффективности маркетинговой акции посредством идентификации товаров и магазинов.

В реляционном моделировании при разрешении отношений "многие ко многим" называются **связывающими таблицами** (associative tables). Таблицы *фактов* охвата событий управляют исключениями, когда некоторая сущность одной таблицы связана с некоторыми сущностями одной или нескольких других таблиц.

*Таблица-мост* (Bridge table) была определена в этой лекции выше. Напомним, что *таблица-мост* является промежуточной таблицей, которая несет на себе функцию управления ситуацией, связанной с реализацией многозначных функциональных зависимостей в данных.

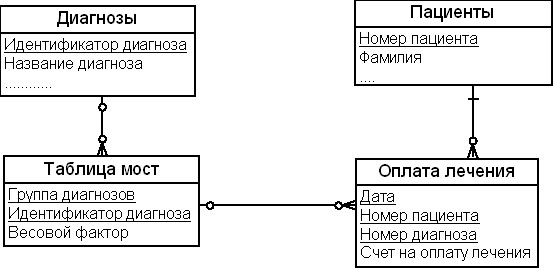
Рассмотрим пример на [рис. 9.25](https://intuit.ru/studies/courses/599/455/lecture/10167?page=7#image.9.25), где представлены два *измерения*: "Пациенты" и "Диагнозы" – и таблица *фактов* "Оплата лечения", которая содержит данные о плате за лечение, полученной с выставленного счета.



**Рис. 9.25.**Схема данных для учета оплаты лечения

*Гранулированность* таблицы *фактов* определена так: на одного пациента существует одна запись о плате за лечение с установленным диагнозом. Допустим, что лечение семьи, переболевшей ОРВИ, оплачивается с одного счета. Чтобы не нарушать определение *гранулированности* для таких случаев, мы можем использовать *таблицу-мост*, как показано на [рис. 9.26](https://intuit.ru/studies/courses/599/455/lecture/10167?page=7#image.9.26).

"Весовой множитель" — важный столбец в *таблице-мосте*. Мы назначаем числовой весовой множитель каждому пациенту для каждого семейного номера счета так, что сумма всех весовых множителей, принадлежащих номеру счета, точно 1. Весовой множитель — это просто способ распределить числовые *аддитивные факты* между пациентами, которые присутствуют в таблице *измерений* "Пациенты".



**Рис. 9.26.**Использование таблицы-моста для разрешения отношения "многие ко многим"

### Резюме

В настоящей лекции мы определили и рассмотрели основные понятия и элементы *многомерной модели*.

К основным понятиям *многомерной модели* относятся *факты*, *измерения*, параметры (метрики) и *атрибуты*, которые хранятся в таблицах *фактов* и таблицах *измерений*.

В *многомерной модели* различают три основных вида таблиц *фактов*:

* *транзакционные таблицы фактов* ;
* *таблицы фактов периодических моментальных снимков* ;
* *таблица фактов кумулятивных моментальных снимков*.

В *многомерной модели* различают следующие основные таблицы *измерений*:

* *вырожденные измерения* ;
* *медленно меняющиеся измерения* ;
* *быстро меняющиеся измерения* ;
* таблицы *иерархий измерений*.

В процессе *многомерного моделирования* используется ряд вспомогательных таблиц, таких как *таблицы-мосты* и *таблицы фактов без метрик*.

Для представления *многомерной модели* используются следующие основные схемы:

* *схема "звезда"* ;
* *схема "снежинка"* ;
* схема с несколькими таблицами *фактов*.

В последующих лекциях мы будем изучать, как применяются рассмотренные понятия в построении моделей данных ХД.