Метод многомерного моделирования

Распределённые базы данных и хранилища данных

Основные понятия

- Многомерное моделирование – это метод моделирования и визуализации данных как множества числовых или лингвистических показателей или параметров (measures), которые описывают общие аспекты деятельности организации.
- Многомерная модель (Dimensional model) ориентирована в первую очередь на выполнение сложных запросов к базе данных.

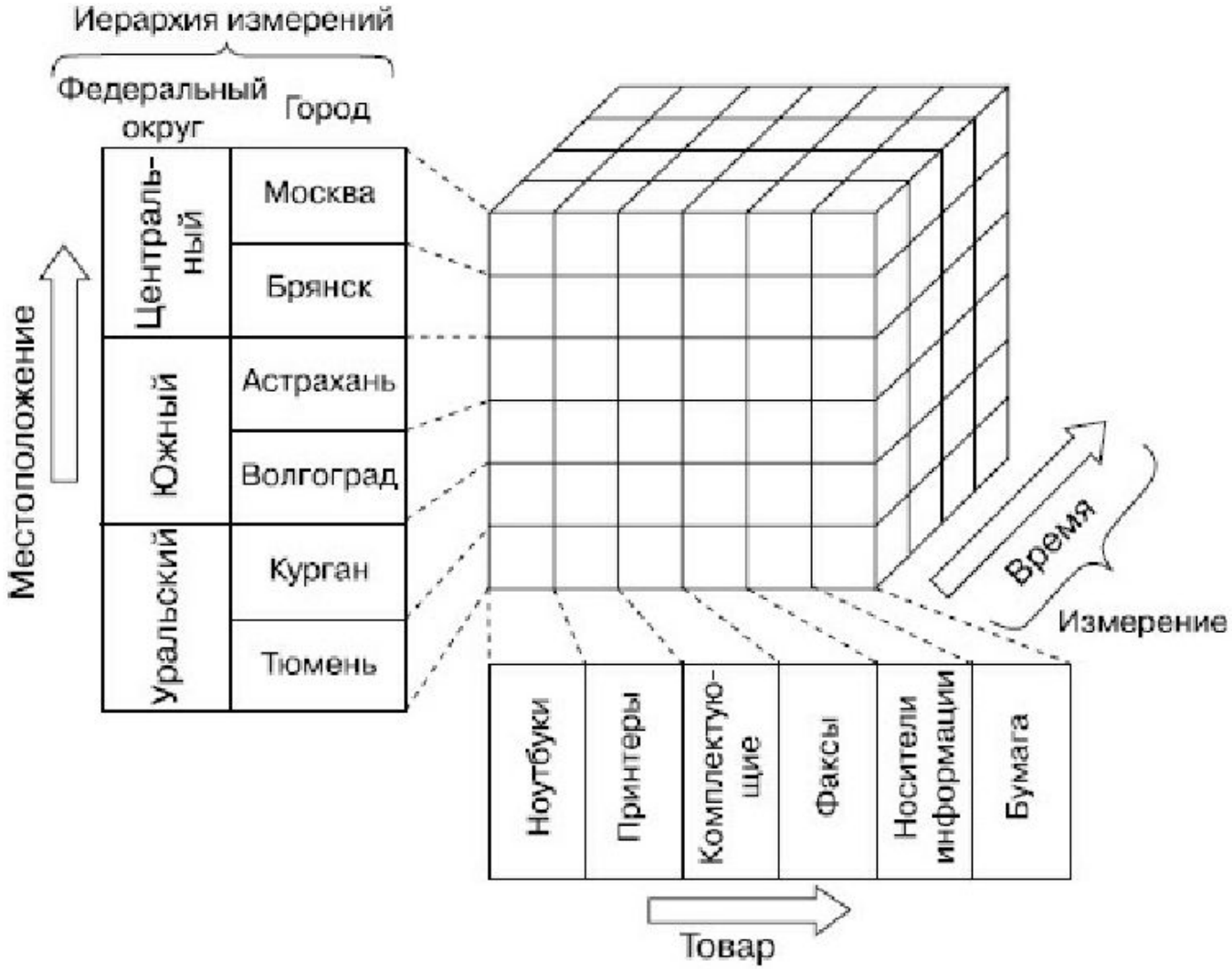
Основные понятия

- 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.
- Атрибут (Attribute) – это описание характеристики реального объекта предметной области.
- Измерение (dimension) – это интерпретация факта с некоторой точки зрения в реальном мире. Обычно измерения представляются как оси многомерного пространства, точками которого являются связанные с ними факты. Измерения задаются перечислением своих элементов (members). Элемент измерения (dimensional member) – уникальное имя или идентификатор (лингвистическая переменная), используемая для определения позиции элемента.

Основные понятия

- Иерархия (hierarchy) – группировка объектов одного измерения в объекты более высокого уровня. Иерархия целиком основывается на одном измерении и формируется из уровней (hierarchy levels).
- Параметр, метрика или показатель (measure) – это числовая характеристика факта, который определяет эффективность деятельности или бизнес-действия организации с точки зрения измерения.
- Гранулированность (Granularity) – это уровень детализации данных, сохраняемых в хранилище данных.

Куб данных


Иерархия измерений
Федеральный округ
Город
Центральный
Южный
Уральский
Москва
Брянск
Астрахань
Волгоград
Курган
Тюмень
Местоположение
Время
Измерение
Ноутбуки
Принтеры
Комплектующие
Факсы
Носители информации
Бумага
Товар

Свертывание и развертывание данных

- Развертка (drill down) и свертка (drill up) являются операциями перемещения вниз и вверх по уровням иерархии измерения.

	20.03	20.04	20.05
Центральный	152500	241400	212600
Южный	50200	53600	51400

Развертка (drill down) по измерению «Регион»

	20.03	20.04	20.05
Москва	131000	220000	200000
Тверь	21500	21400	12600
Ростов	20100	21600	21400
Волгоград	30100	32000	30000

	год
Центральный	212900
Южный	40700

Свертка (drill up) по измерению «Время»

	1 кв.	2 кв.	3 кв.	4 кв.
Центральный	53600	53150	53000	53150
Южный	12850	12700	2300	12850

Классы фактов

- 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.
- 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.
- 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.
- 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.

Примеры

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

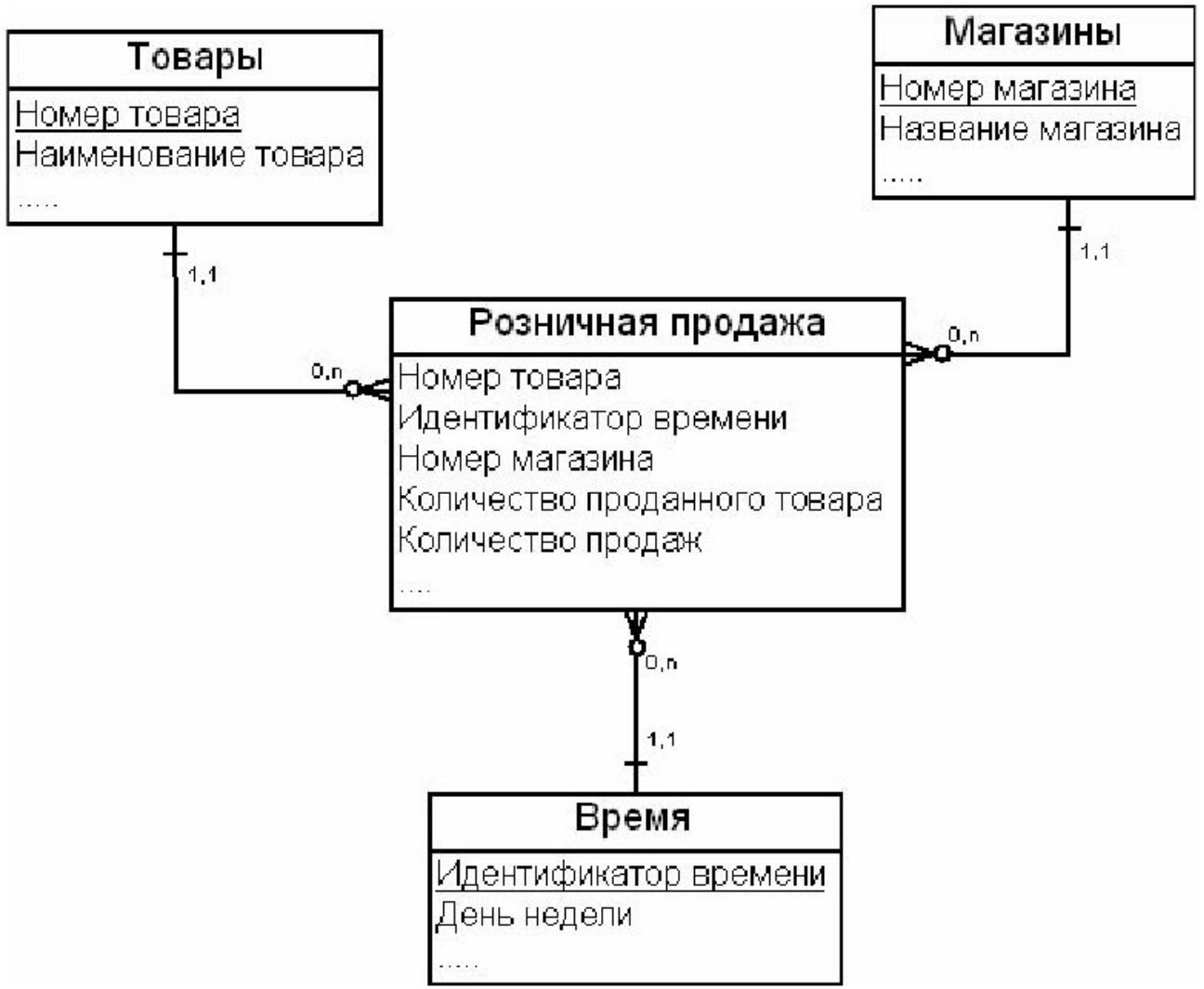
По измерению "Магазин":

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

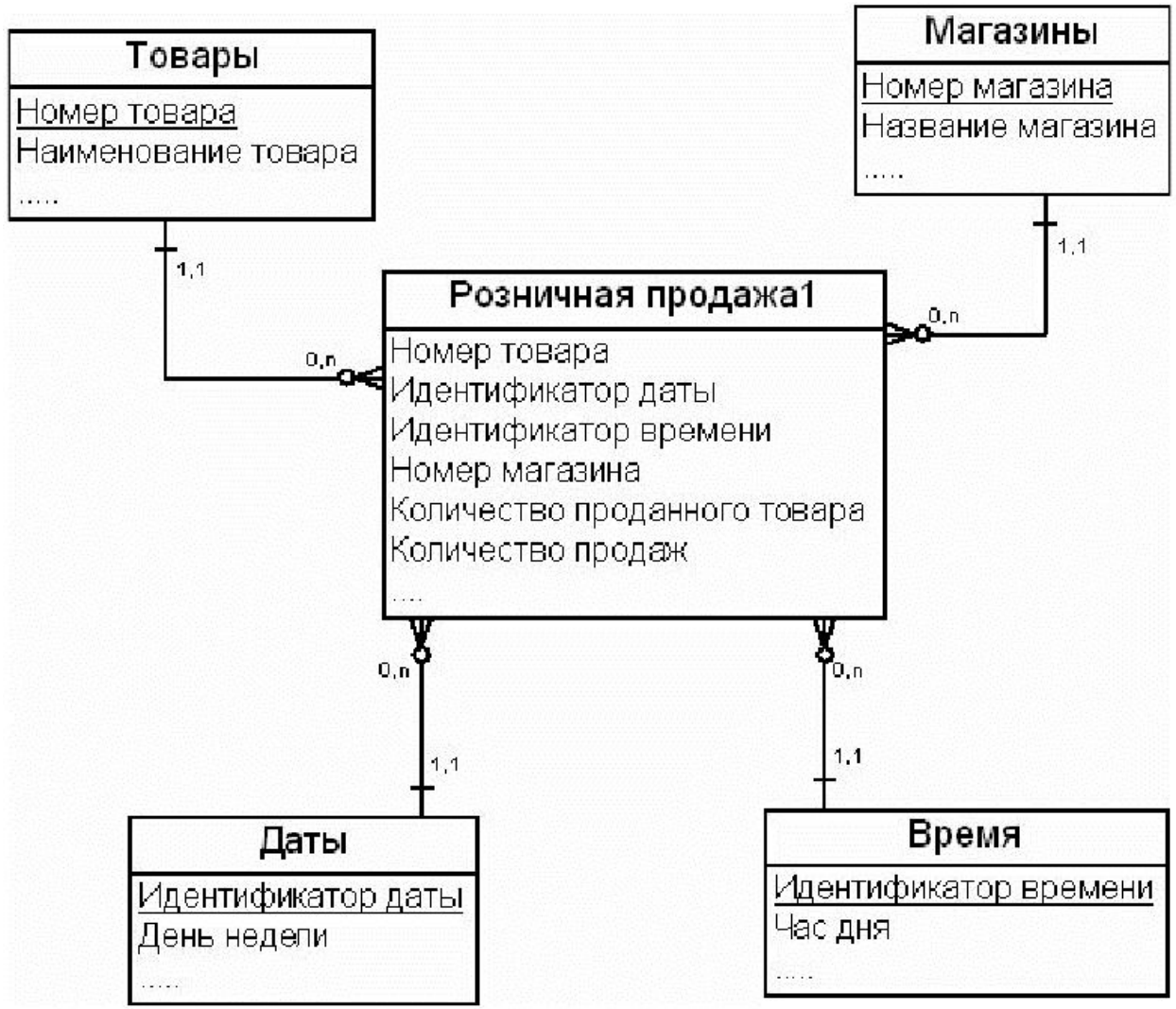
Ключи в таблице фактов

- 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.
- Гранулированность фактов определяет смысл значения факта с точки зрения уровня детализации, связываемой с фактом информации.

Пример с уникальным первичным ключом таблицы фактов (гранулированность фактов – одни сутки)



Пример с уникальным первичным ключом таблицы фактов (гранулированность фактов – один час)



Категории таблиц фактов

- Транзакционная таблица фактов. В такой таблице фактов сохраняют факты, которые фиксируют определенные события (транзакции).
- Таблица фактов периодических моментальных снимков. В такой таблице собирают факты, фиксирующие текущее состояние определенного направления бизнеса.
- Таблица фактов кумулятивных моментальных снимков. В такой таблице собирают факты, фиксирующие некоторое итоговое состояние определенного направления бизнеса на текущий момент времени.

Основные характеристики таблицы фактов

- Таблица фактов содержит числовые параметры (метрики).
- Каждая таблица фактов имеет составной ключ, состоящий из первичных ключей таблиц измерений. Первичный ключ таблицы измерений является внешним ключом в таблице фактов.
- Таблица фактов имеет, как правило, небольшое количество полей.
- 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.

Основные характеристики таблицы измерений

- Таблицы измерений содержат данные о детализации фактов.
- Таблицы измерений содержат описательную информацию о числовых значениях в таблице фактов.
- Как правило, денормализованные таблицы измерений содержат большое количество полей.
- Таблицы измерений содержат обычно значительно меньше строк, чем таблицы фактов.
- Атрибуты таблиц измерений обычно используются при визуализации данных в отчетах и запросах.

Примеры таблицы фактов и таблицы измерений

- Таблица фактов

Продажи
Идентификатор времени
Идентификатор магазина
Идентификатор товара
Суммарная прибыль
Количество покупателей
Количество продаж

- Таблица измерений

Время
Идентификатор времени
День недели
День месяца
Месяц
Квартал
Год
Отчетный период
Флажок праздника

Схемы многомерной модели

- Схема «звезда» (star schema) имеет одну таблицу фактов и несколько таблиц измерений.
- Схема «снежинка» (snowflake schema) имеет одну таблицу фактов и несколько нормализованных таблиц измерений.
- Схема «созвездие» (fact constellation schema) имеет несколько таблиц фактов.

Схема «звезда»

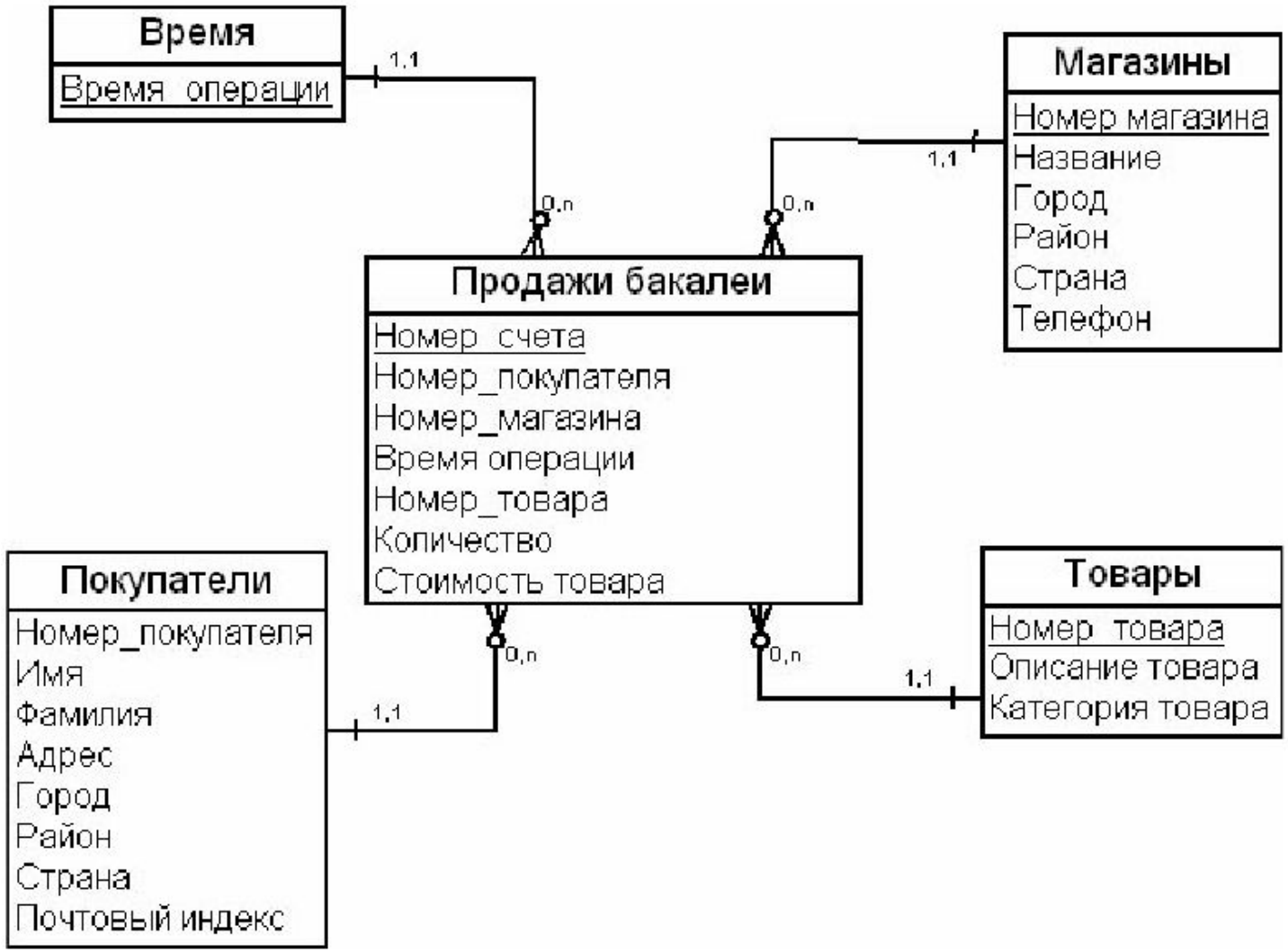


Схема «снежинка»

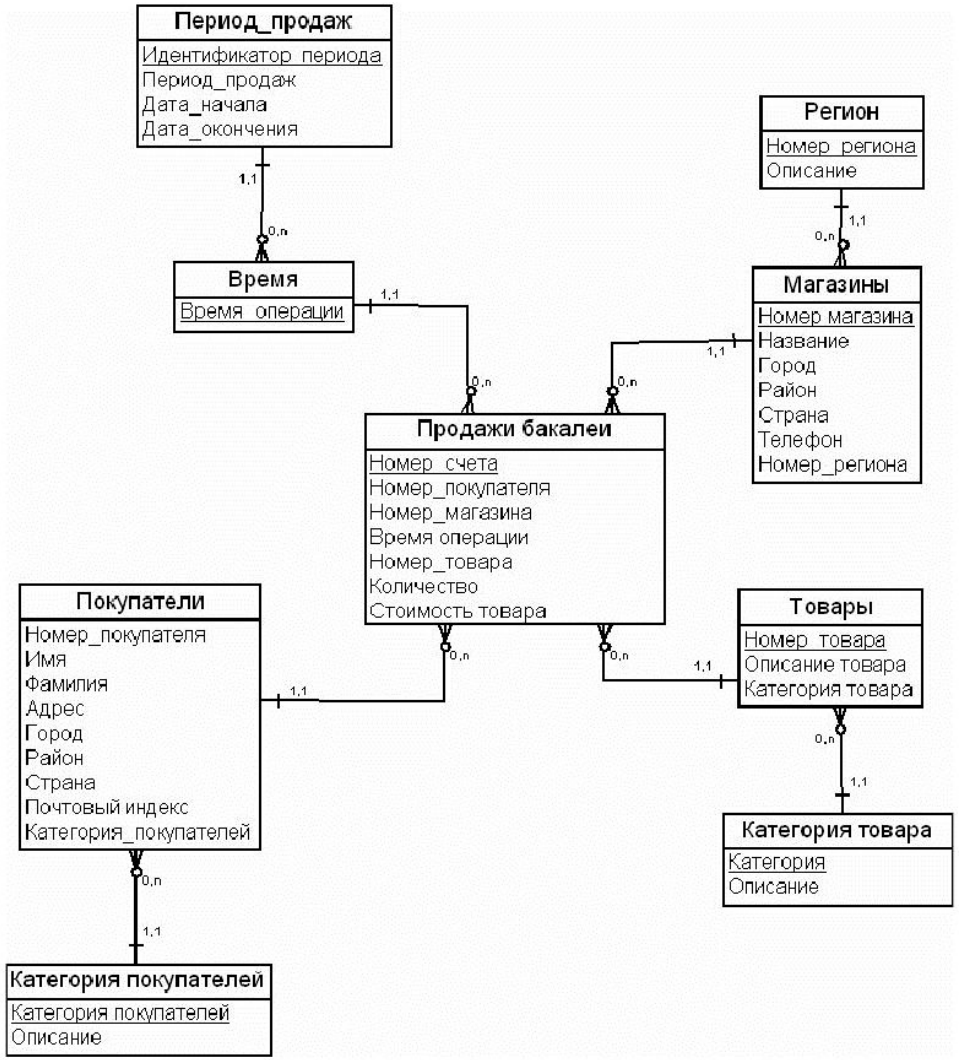


Схема «снежинка»

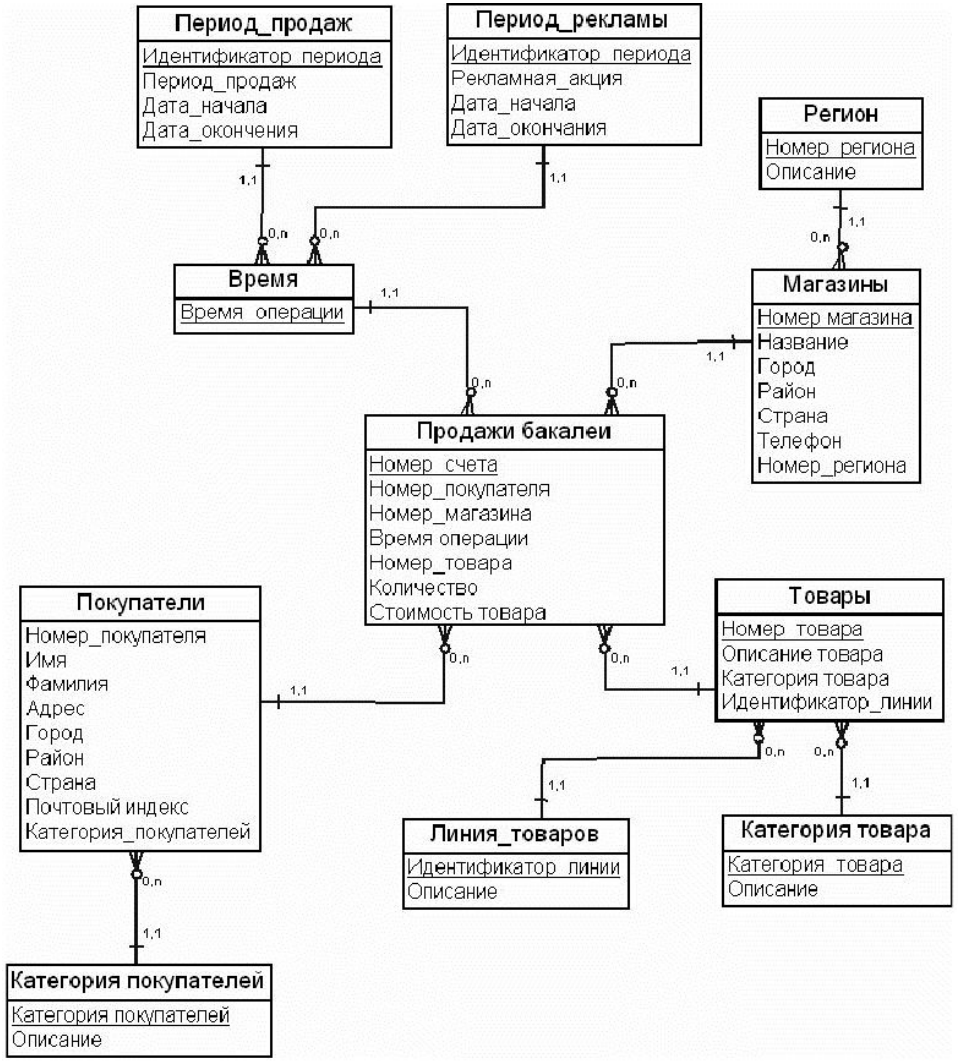
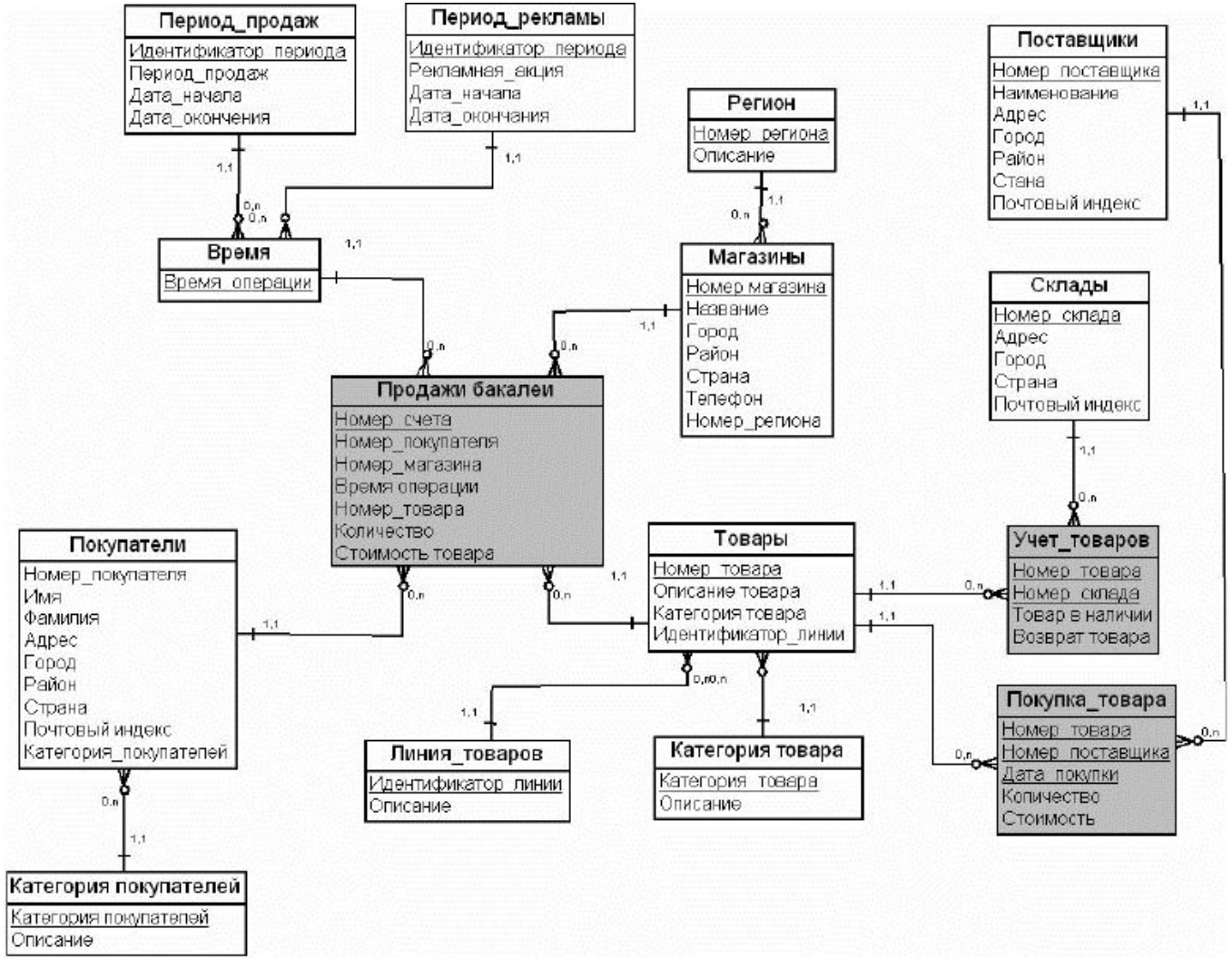


Схема «созвездие»

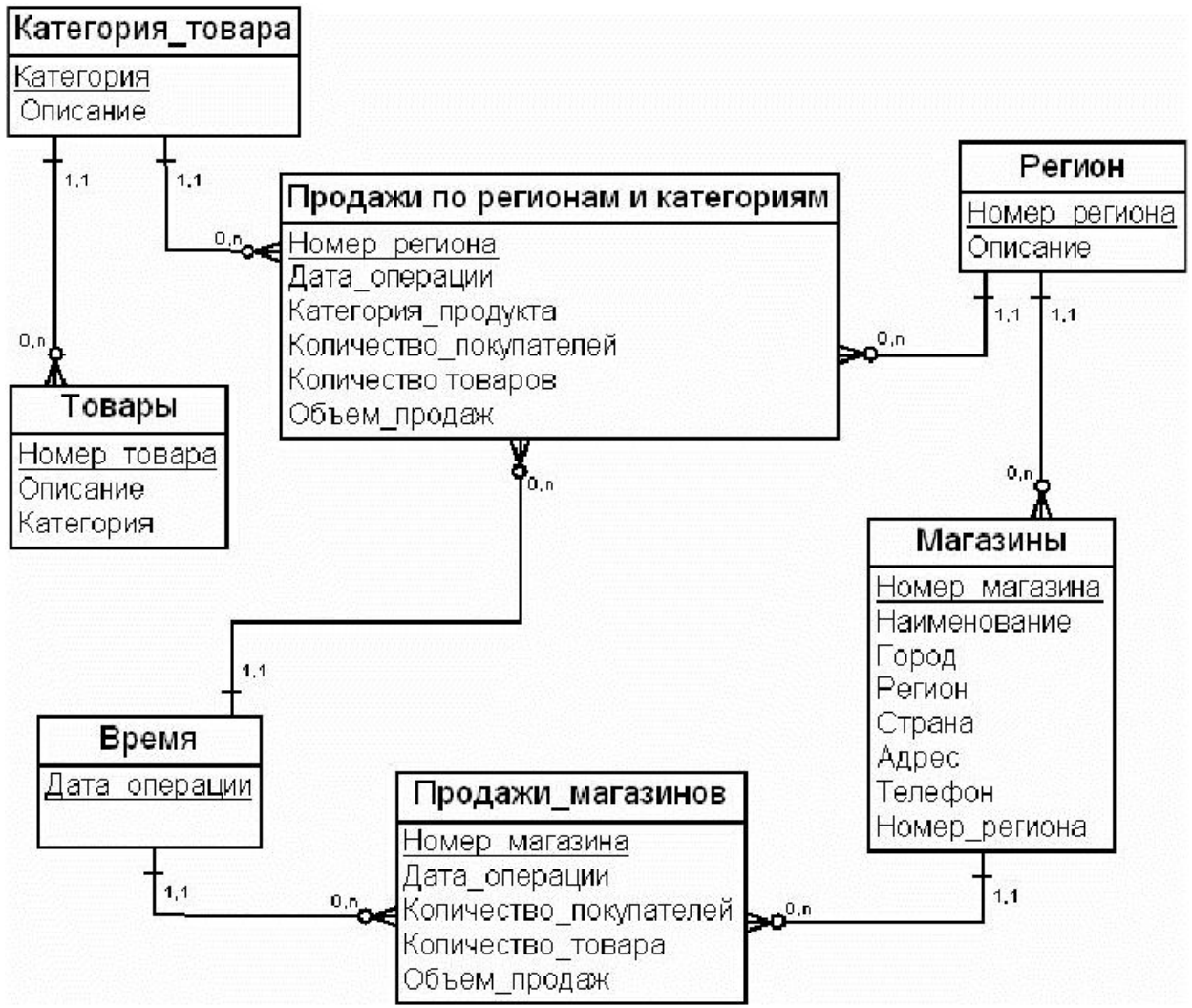


Моделирование таблиц фактов

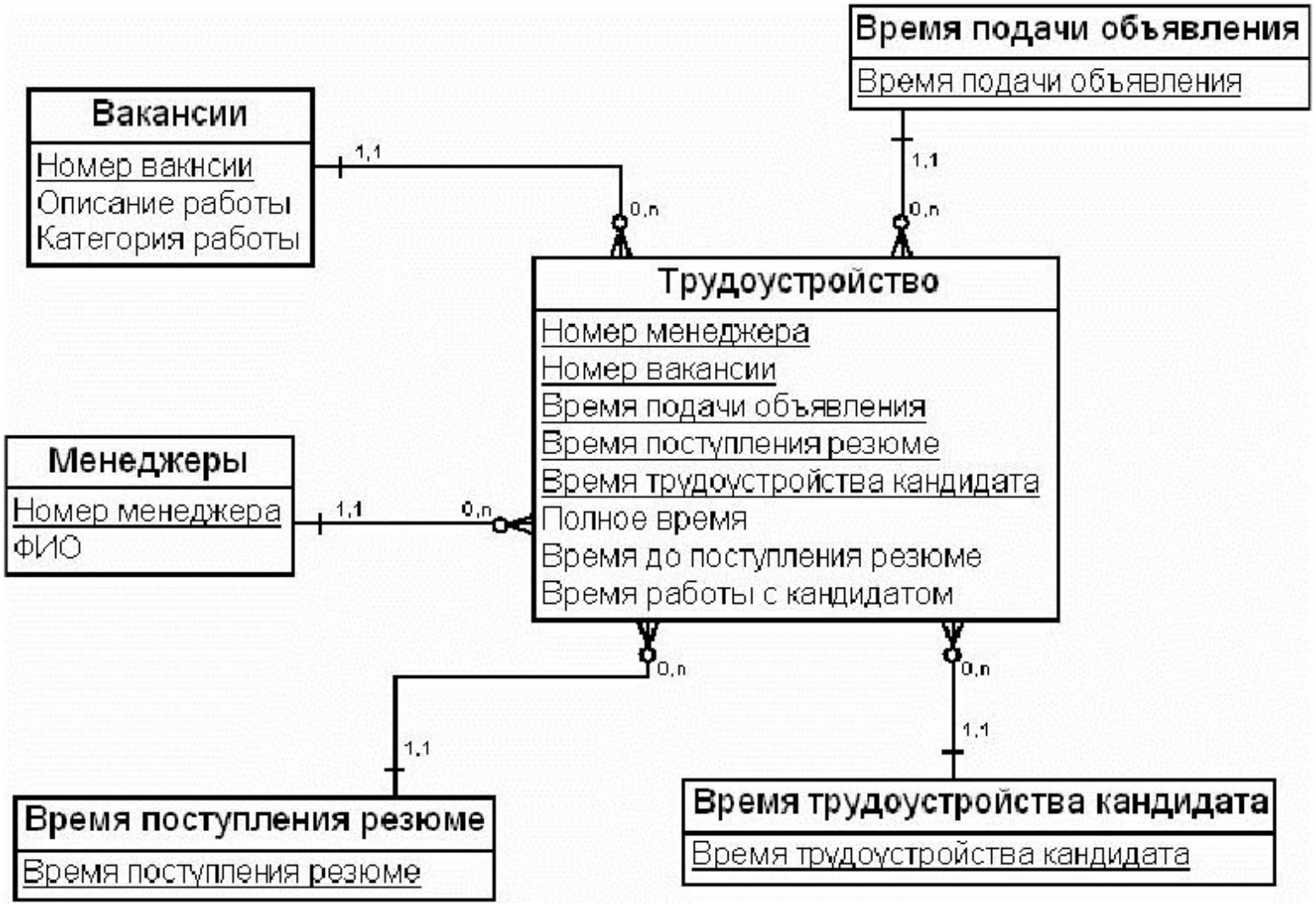
- Агрегатами являются суммы значений параметров или статистические функции от значений параметров, взятые на определенном уровне детализации (гранулированности).
- Таблицей агрегатов фактов (Aggregate fact table) называется таблица фактов, которая содержит агрегаты некоторых фактов модели.
- Обычно в хранилище данных используют два типа таблиц агрегатов фактов:
  - со степенью детализации на уровне периодического снимка данных, представляющего промежуток времени заданной продолжительности (таблица фактов периодических моментальных снимков);
  - 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).

Пример таблицы агрегатов фактов периодических моментальных снимков

Пример таблицы агрегатов фактов периодических моментальных снимков



Пример таблицы агрегатов фактов кумулятивных моментальных снимков



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 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

Моделирование таблиц измерений

- 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.
- Типы 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.

Медленно меняющиеся измерения (тип 1)

- Старое значение атрибута меняется на новое значение.

Табельный номер	Фамилия	Имя	Семейное положение
332201	Иванова	Анна	Не замужем

<===== Замужем

Медленно меняющиеся измерения (тип 2)

- Создается новая запись в таблице измерения с новым суррогатным ключом.

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

Медленно меняющиеся измерения (тип 3)

- Создаются новые поля в таблице измерения.

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

Пример медленно меняющегося измерения (тип 2)

Пример медленно меняющегося измерения (тип 3)

Схема принятия решения при выборе типа медленно меняющегося измерения

ЕСЛИ требуется сохранять историю измерения, ТО следует выбрать тип 2

В ПРОТИВНОМ СЛУЧАЕ

ЕСЛИ необходимо сравнивать текущее значение атрибута с первоначальным или предыдущим, ТО следует выбрать тип 3

В ПРОТИВНОМ СЛУЧАЕ

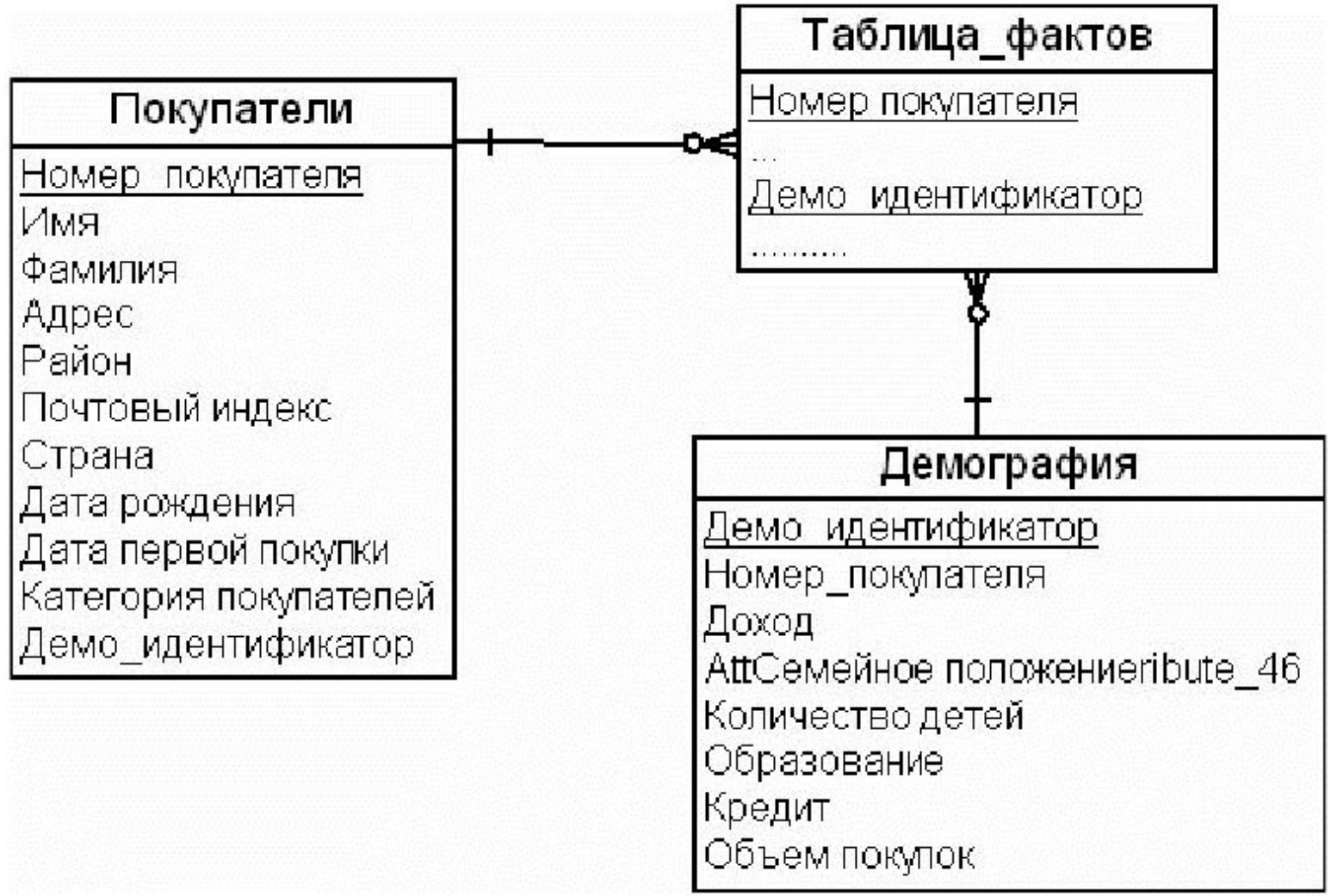
следует выбрать тип 1.

Моделирование таблиц измерений

- 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.

Пример быстро меняющегося измерения

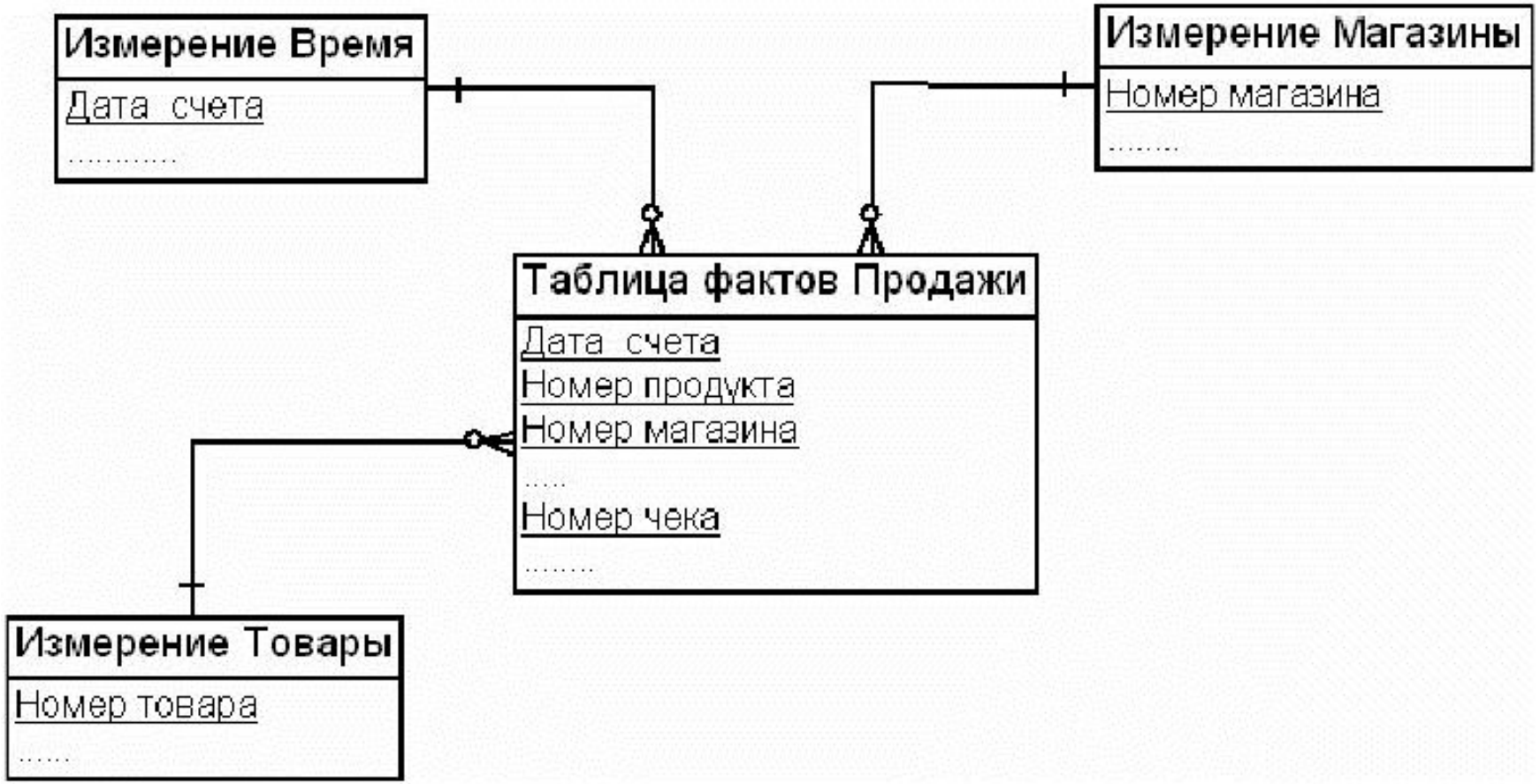
Пример разбиения быстро меняющегося измерения



Моделирование таблиц измерений

- 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.
- Обычно вырожденное измерение представлено атрибутами ключа измерения в таблице фактов без соответствующей таблицы измерений.

Пример вырожденного измерения

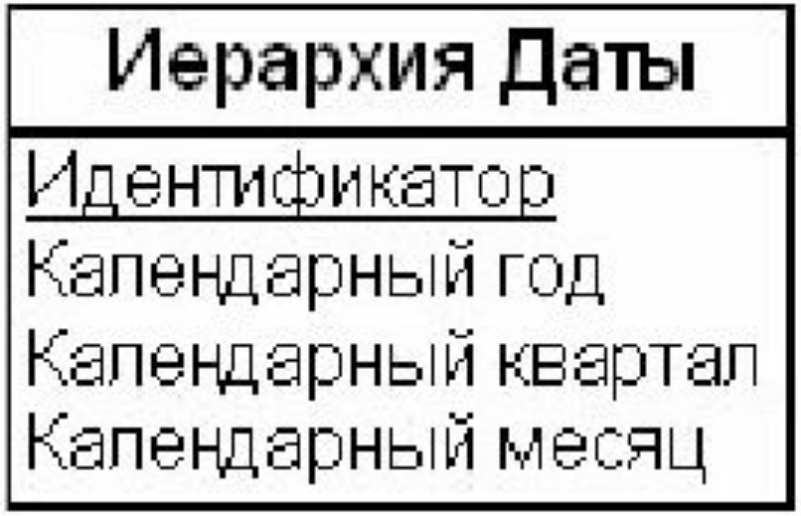


Иерархии измерений

- Иерархией называется взаимосвязанный набор отношений «многие к одному», состоящий из последовательности уровней.
- В многомерном моделировании различают три типа иерархий:
  - сбалансированные иерархии (Balanced hierarchy);
  - несбалансированные иерархии (Unbalanced hierarchy);
  - иерархии с пропущенными уровнями (Ragged hierarchy).

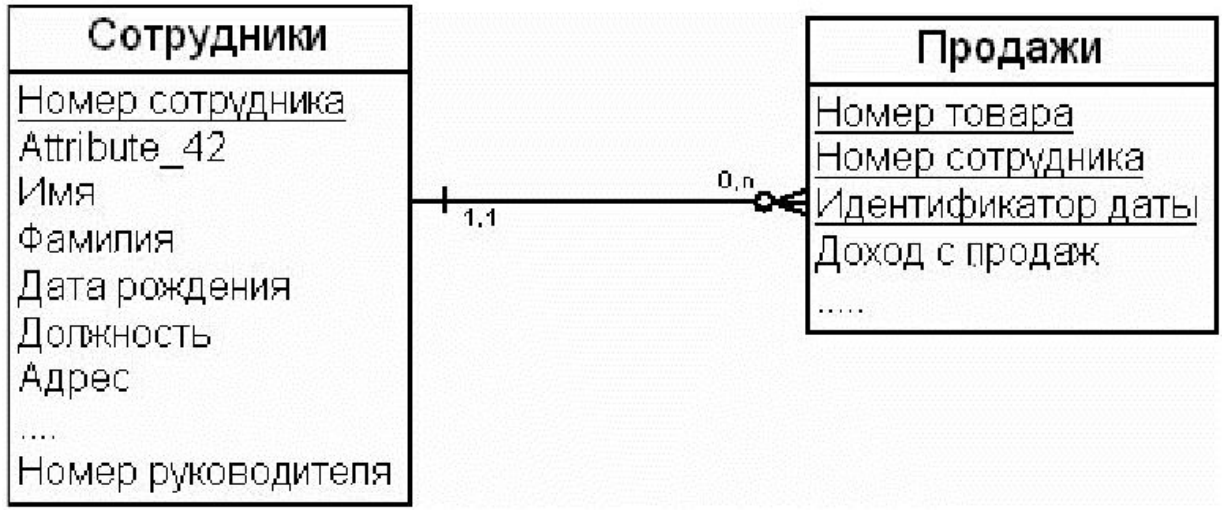
Сбалансированная иерархия

- Сбалансированная иерархия – это иерархия, в которой все ветви измерения имеют одно и то же количество уровней.
- Сбалансированная иерархия состоит из фиксированного числа уровней.



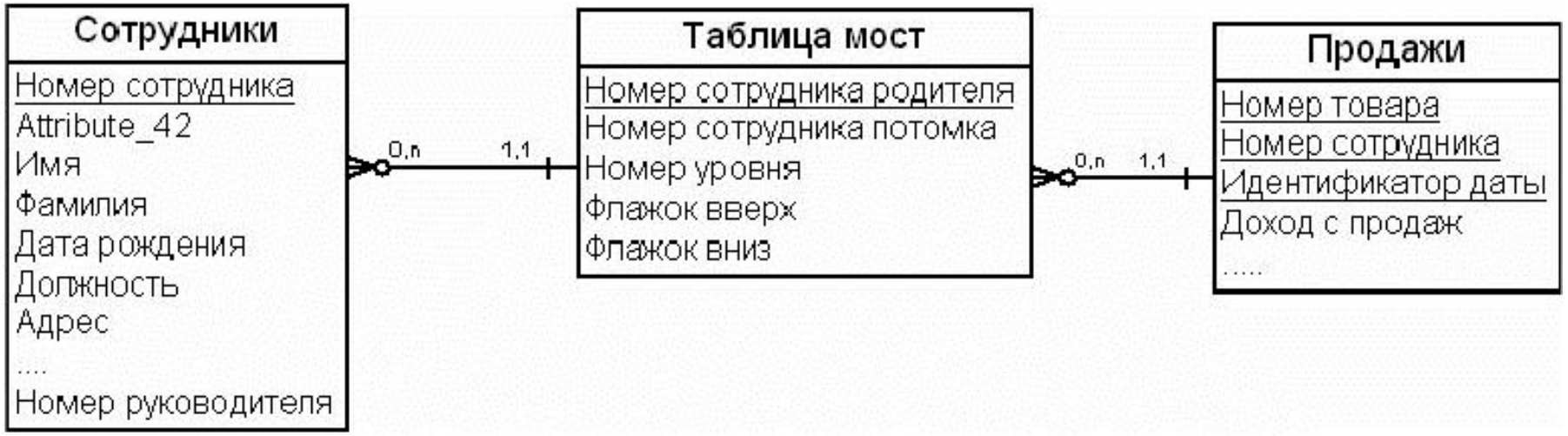
Несбалансированная иерархия

- Несбалансированная иерархия – это иерархия, в которой все ветви измерения имеют различное число уровней.
- Рекурсивный указатель (recursive pointer) – ключ сущности потомка в сущности родителя.



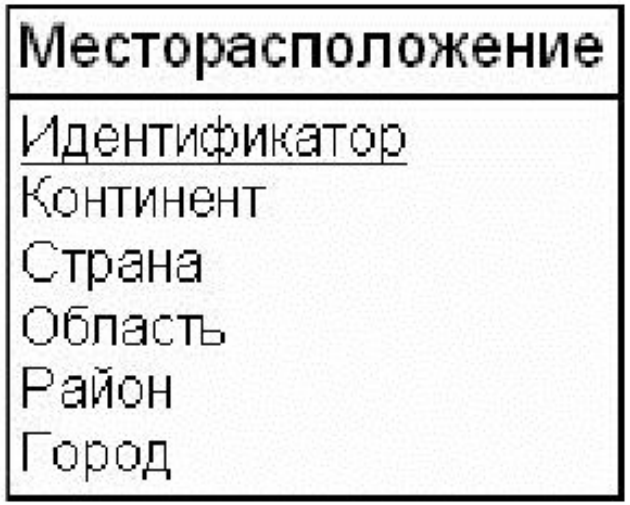
Несбалансированная иерархия

- Таблица-мост (bridge table) – вспомогательная таблица, которая предназначена упростить работу с рекурсивными отношениями, отношениями «многие ко многим», отношениями типа иерархии при использовании реляционной модели данных.



Иерархия с пропущенными уровнями

- Иерархия с пропущенными уровнями – это иерархия, в которой допускается отсутствие одного из уровней при заполнении ее данными.



Отношение «многие ко многим» в измерениях

- Таблицы измерений могут находиться в отношении «многие ко многим» между собой.
- Отношение «многие ко многим» может существовать между:
  - таблицей измерения и таблицей фактов;
  - между таблицами измерений.
- В многомерном моделировании хранилища данных 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);
  - таблица-мост (bridge table).

Таблица фактов без метрик

- Таблица фактов без метрик – это 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».
- Типы таблиц фактов без метрик:
  - таблицы фактов отслеживания событий (event tracking tables);
  - таблицы фактов охвата событий (coverage tables).

Таблица фактов отслеживания событий

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

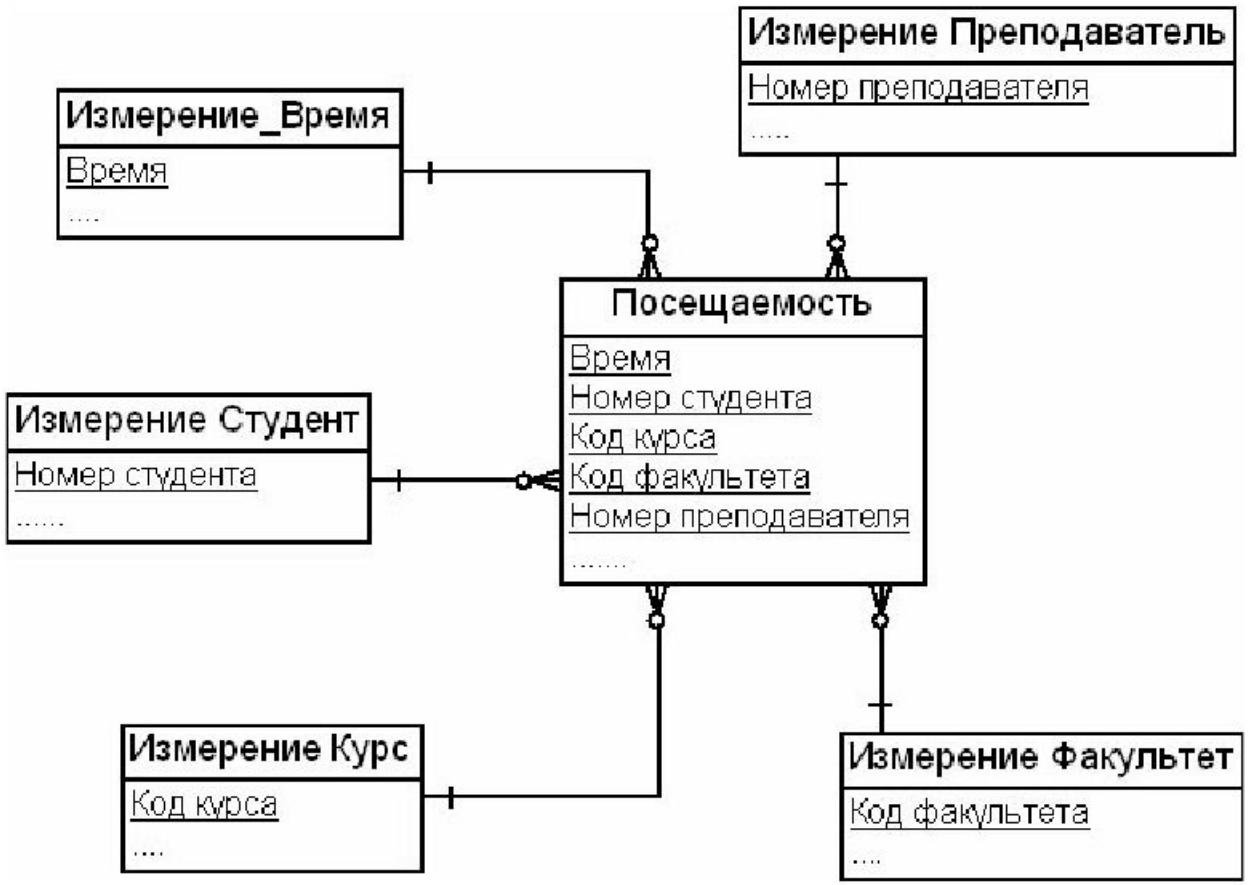


Таблица фактов охвата событий

- Таблица фактов охвата событий содержит описание того, что еще не произошло.

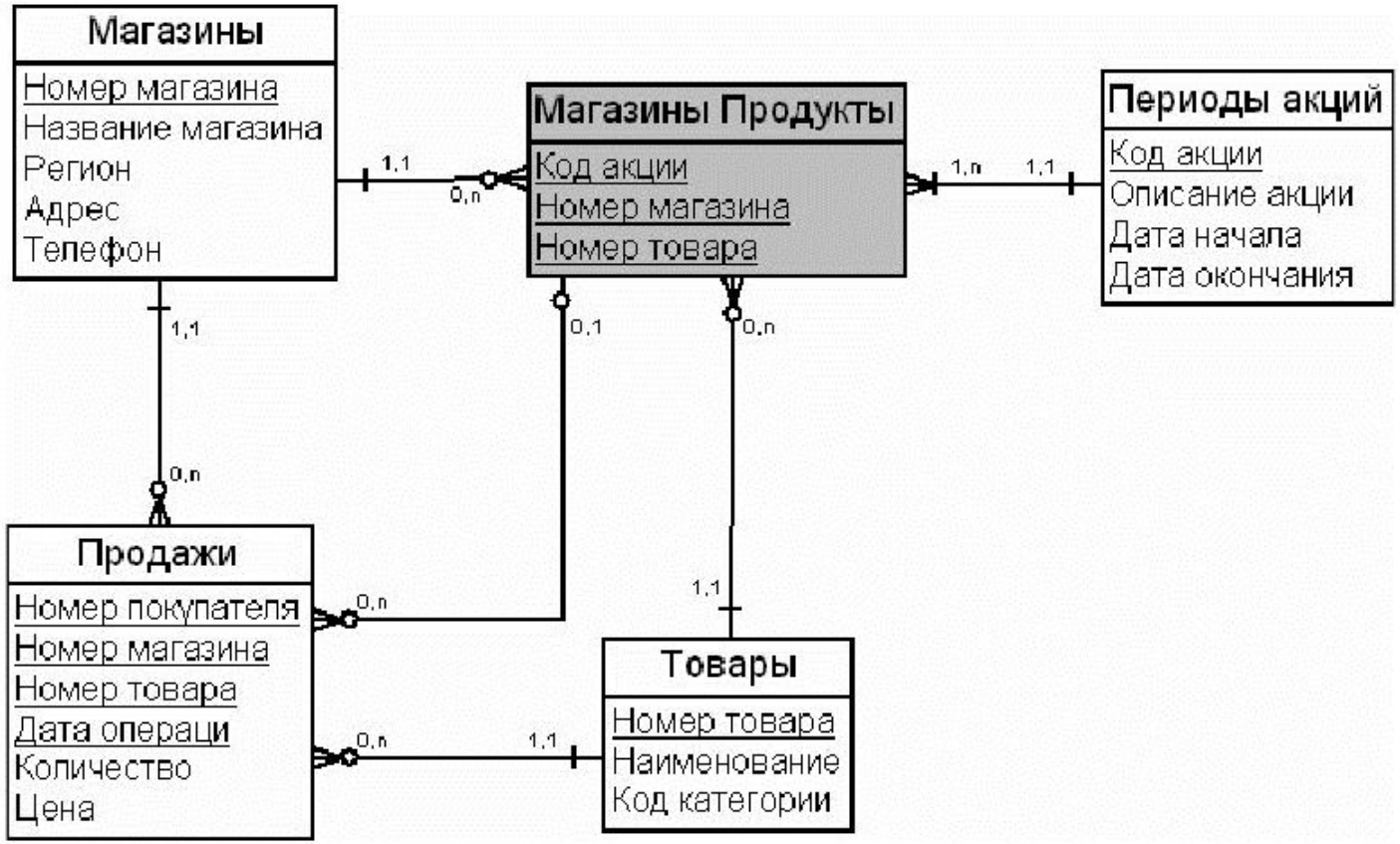


Таблица-мост