Модели данных

A5_Многомерные модели данных

Московский государственный технический университет
имени Н.Э. Баумана

Факультет ИБМ

Июль 2024 года

Москва

Модели реляционных витрин данных ROLAP

- Определение витрины данных
- Многомерная модель данных (гиперкуб)
- Основные элементы гиперкуба (факты, измерения, атрибуты)
- Простые и иерархические измерения (сбалансированные, неровные, несбалансированные, альтернативные иерархии)
- Таксономия многомерной модели данных
- Пример таксономии гиперкуба
- Схема «снежинка» (snowflake)
- Схемы «звезда» (star) и «созвездие» (constellation)

Актуализировать, когда устоится

Витрина данных (Data Mart)

Тематическая база данных, содержащая информацию по отдельным аспектам деятельности организации, предназначенная для обработки средствами бизнес-аналитики.

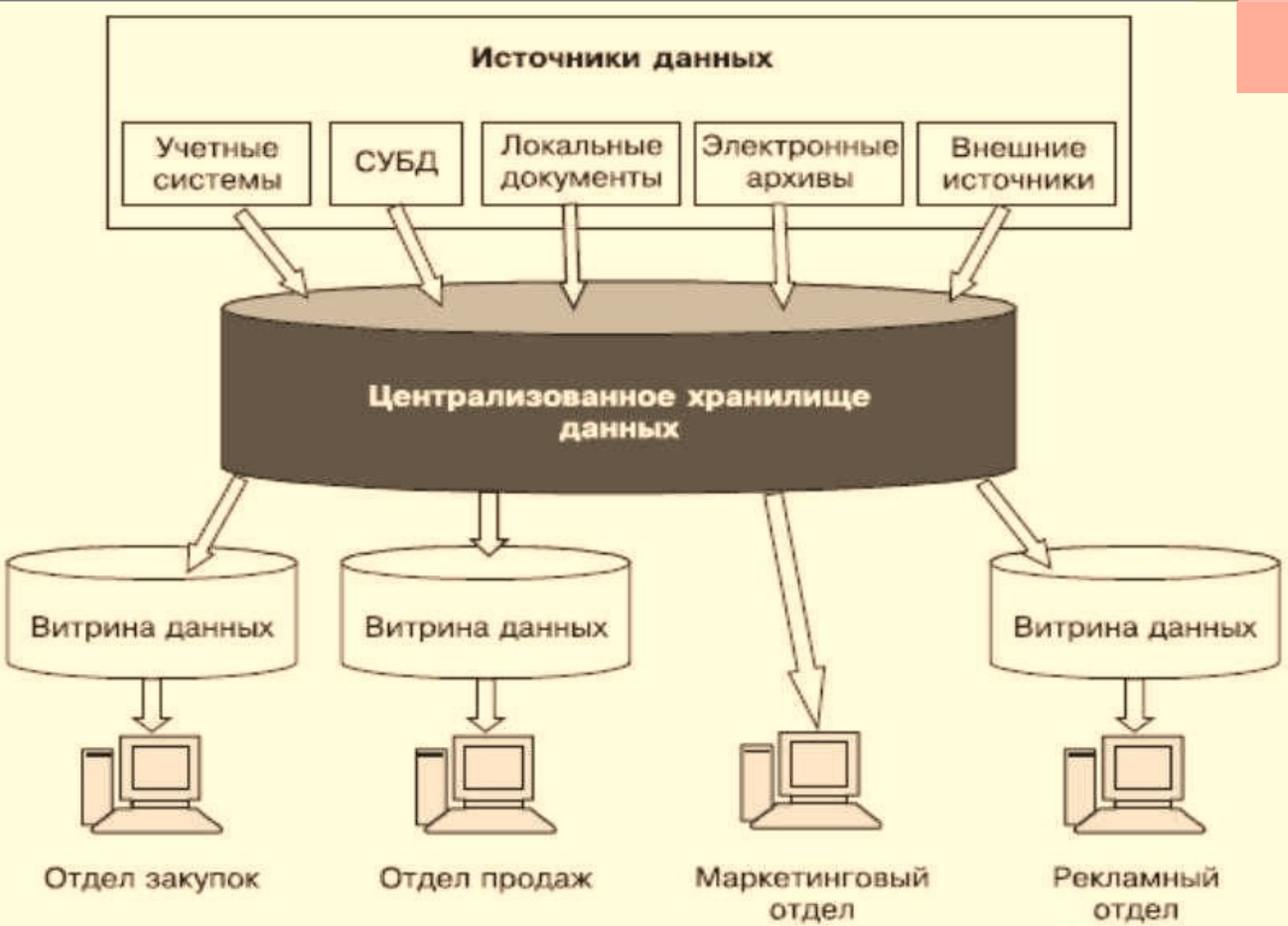
Основные цели выделения витрин:

- приближение данных к конечным пользователям,
- ограничение только необходимыми данными,
- повышение оперативности путём предагрегирования данных,
- ориентация структуры данных на бизнес-аналитику.

Могут быть независимыми от хранилища данных или зависеть и быть частью хранилища данных.

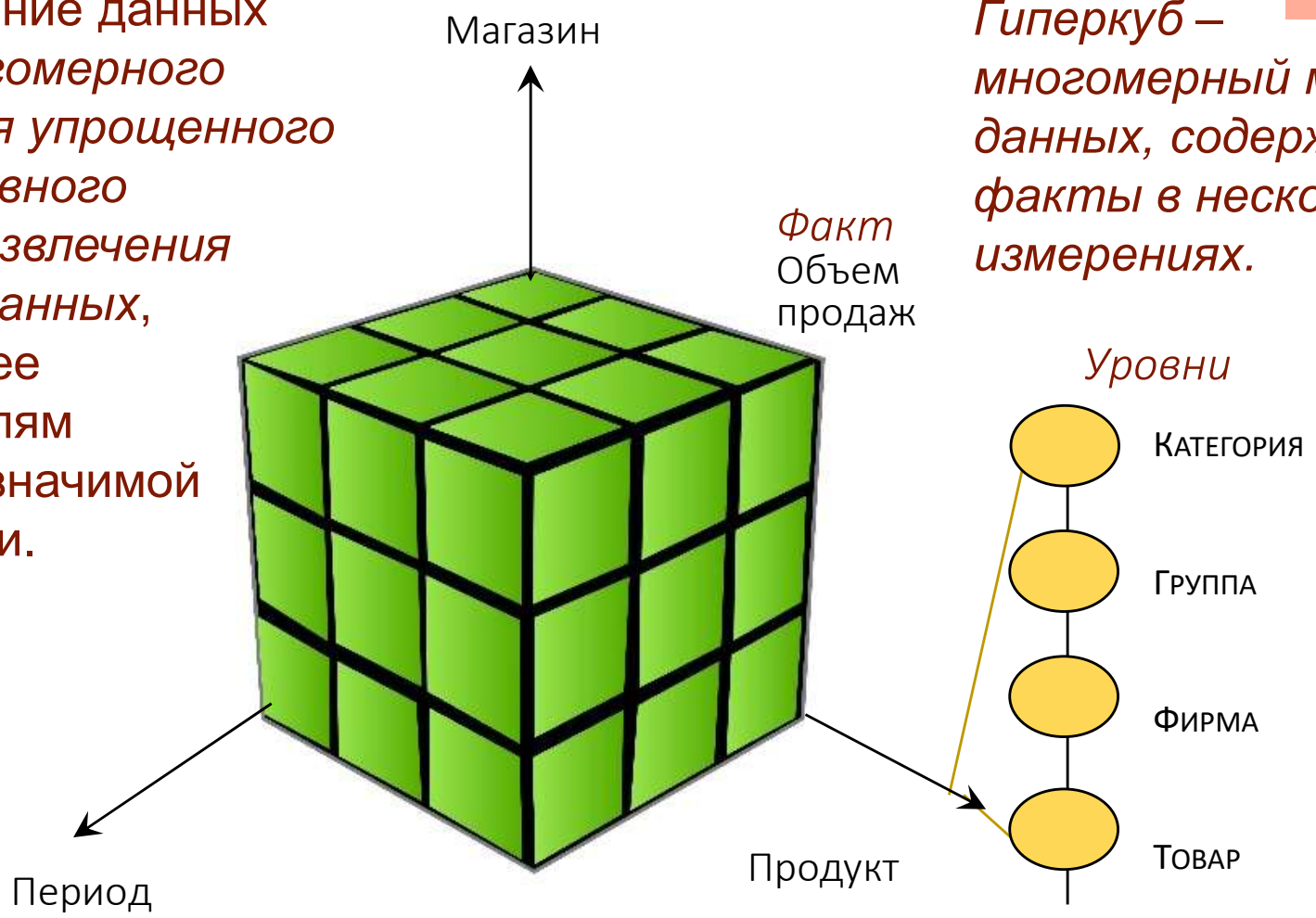
Часто используется многомерная абстракция для моделирования данных, реализации баз данных и обработки данных.

Место витрин данных в аналитических системах



Многомерная модель данных (гиперкуб, факты, измерения)

Представление данных в виде *многомерного массива для упрощенного и эффективного хранения, извлечения и анализа данных*, облегчающее пользователям получение значимой информации.



Гиперкуб – многомерный массив данных, содержащий факты в нескольких измерениях.

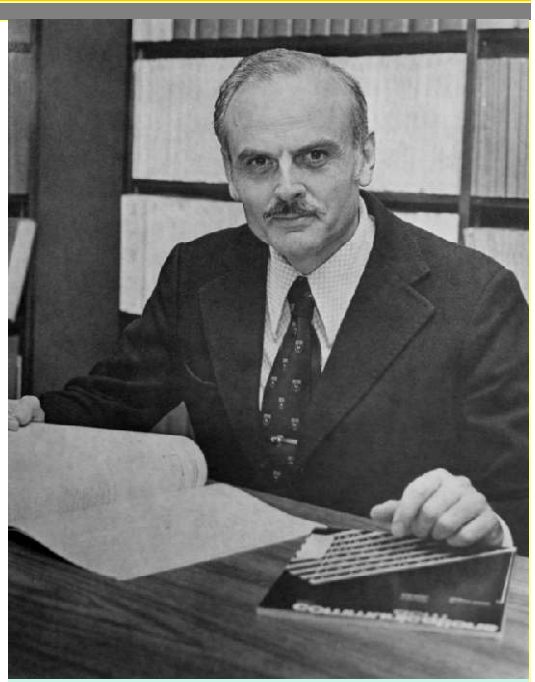
Измерения могут быть *простыми,* а также *иерархическими*

Обобщение реляционной модели от таблиц к массивам

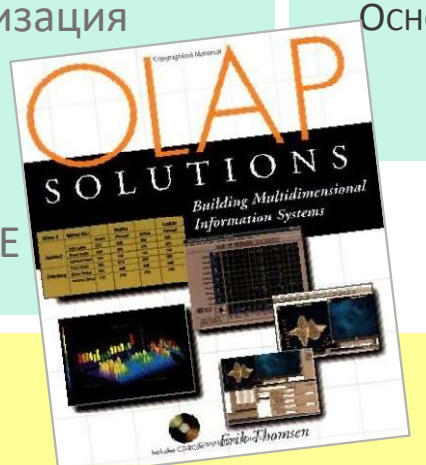
Интерактивная аналитическая обработка (On Line Analytical Processing, OLAP)

Технология анализа данных в различных разрезах и с разной степенью детальности, осуществляемого бизнес–пользователями в интерактивном режиме в терминах своей предметной области.

- Описательный и разведочный анализ
  Многомерный анализ данных (детализация и укрупнение)
- Самообслуживание пользователей в бизнес-терминах
  Привлечение опыта и интуиции пользователя
- «Ручная раскопка» данных (детализация и укрупнение)
  Интерактивные запросы и отчёты, расчёты «на лету»
  Инструменты NO CODE / LOW CODE

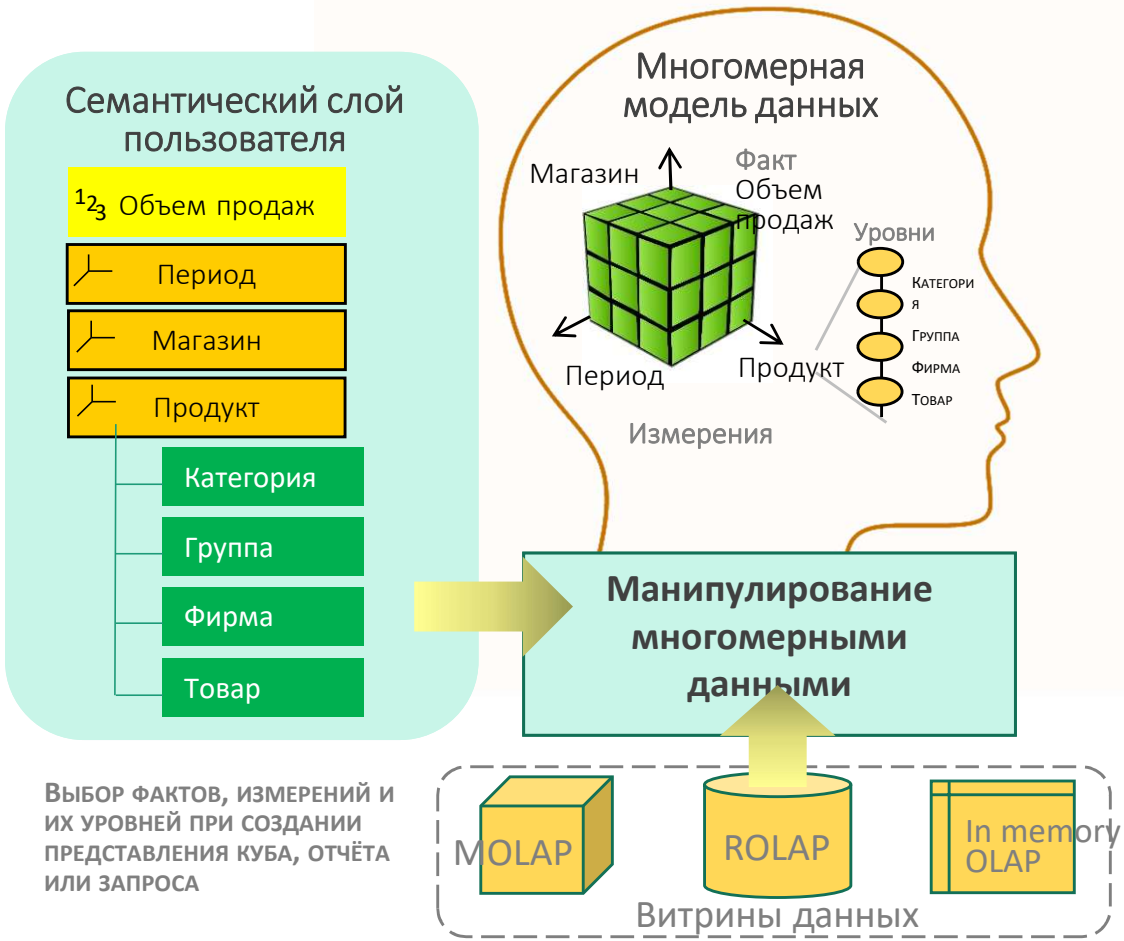


EDWARD CODD
IBM 1993
ОСНОВАТЕЛЬ РЕЛЯЦИОННОЙ МОДЕЛИ ДАННЫХ



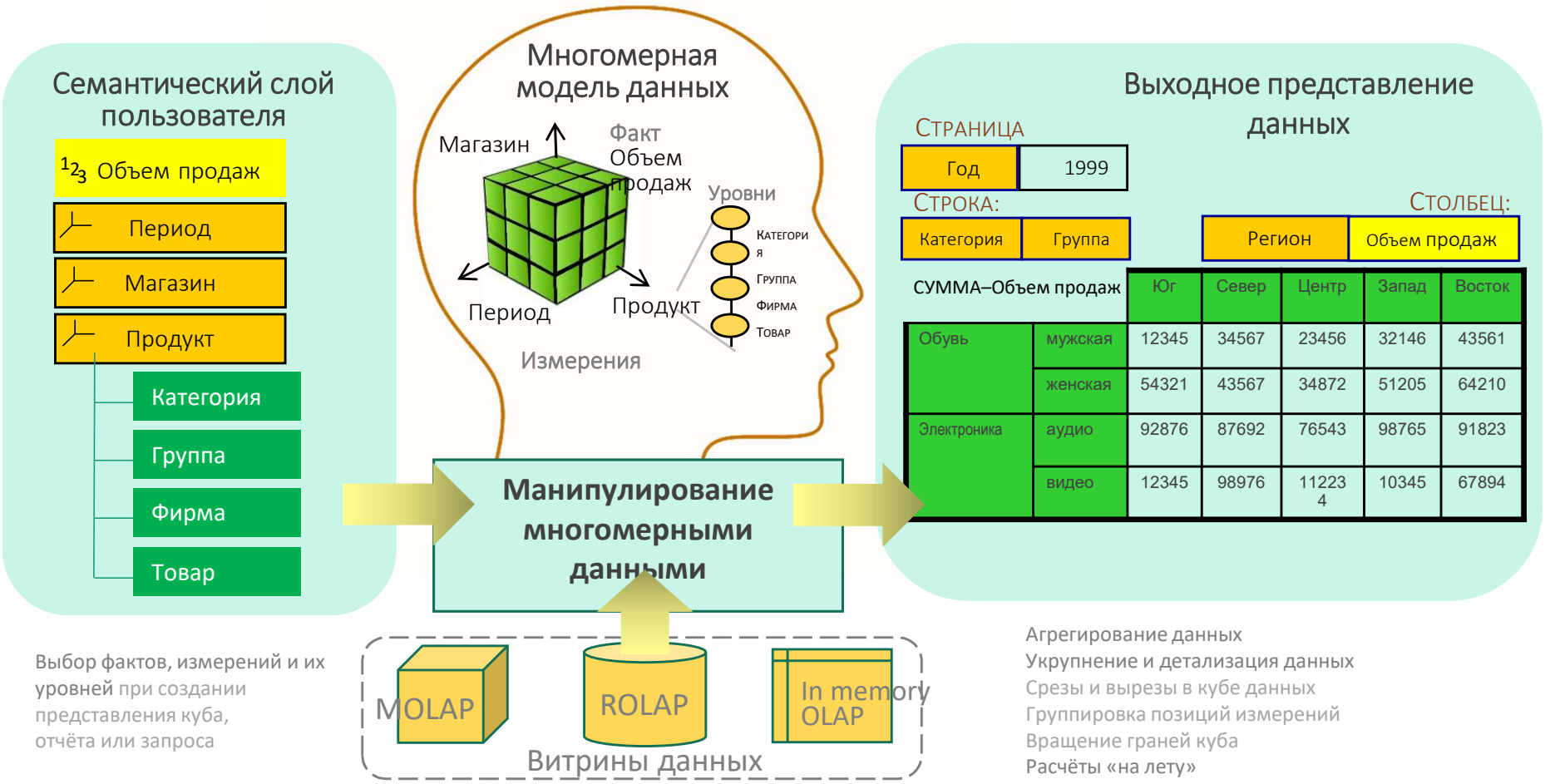
Многомерные свойства OLAP

Многомерные свойства OLAP



Выбор фактов, измерений и их уровней при создании представления куба, отчёта или запроса

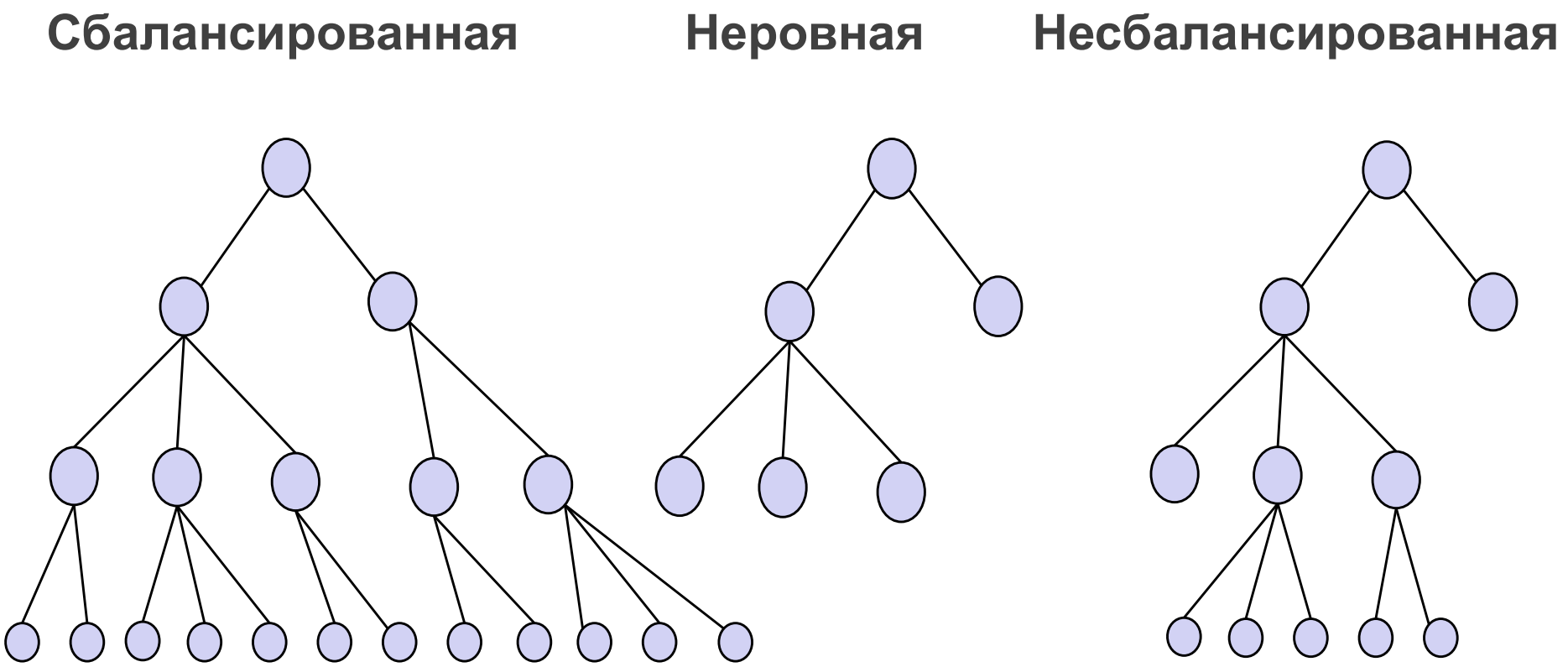
Многомерные свойства OLAP



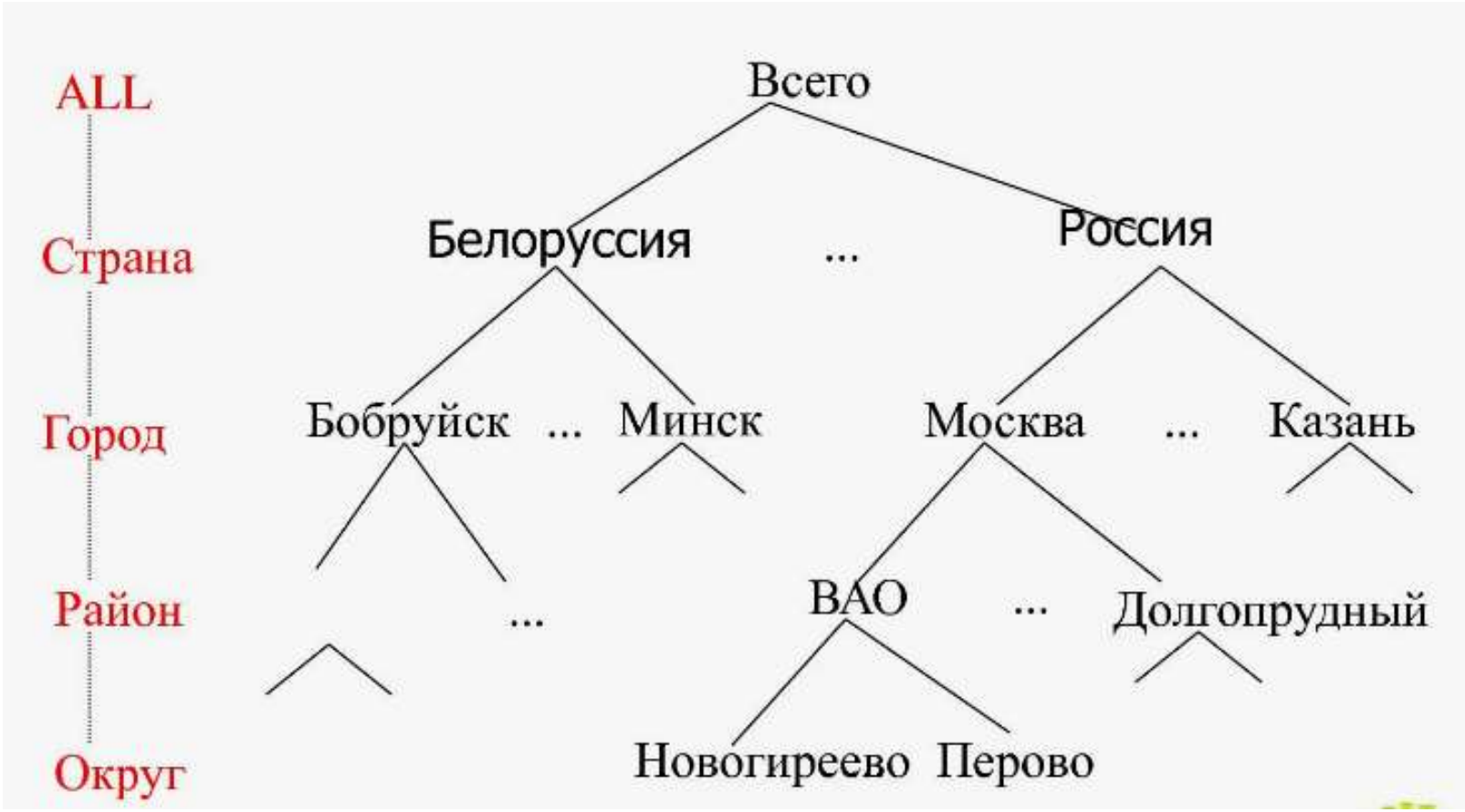
Выбор фактов, измерений и их уровней при создании представления куба, отчёта или запроса

Агрегирование данных
Укрупнение и детализация данных
Срезы и вырезы в кубе данных
Группировка позиций измерений
Вращение граней куба
Расчёты «на лету»

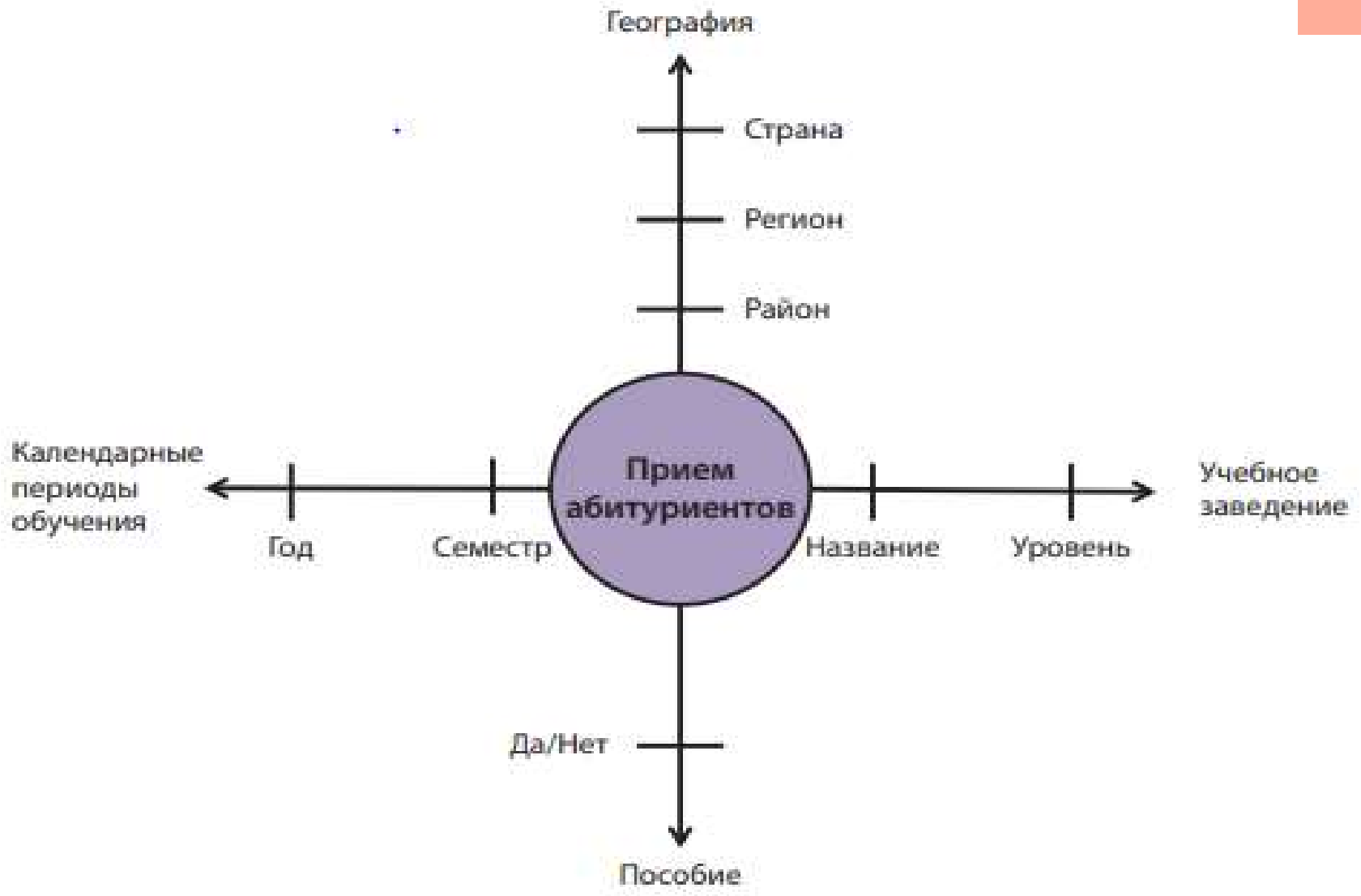
Виды иерархий измерений



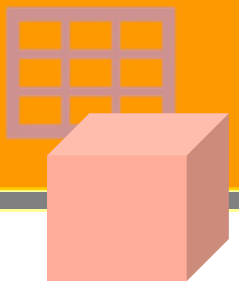
Пример иерархического измерения

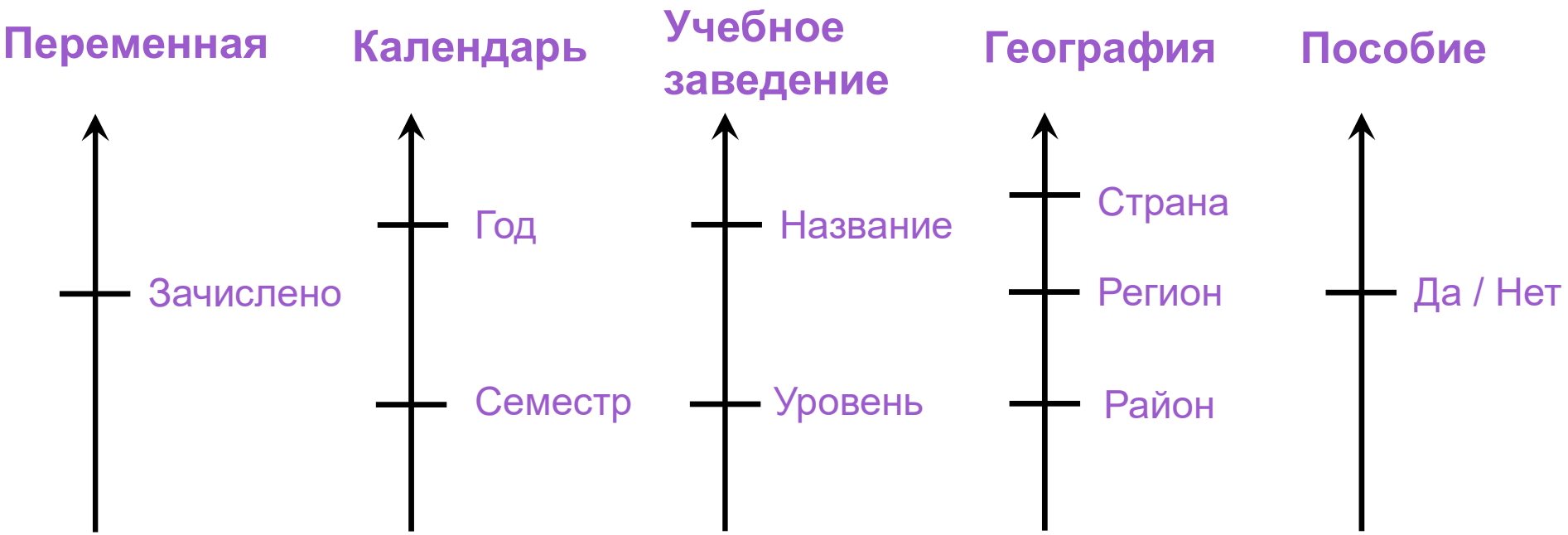


Осевая нотация для представления многомерной модели данных

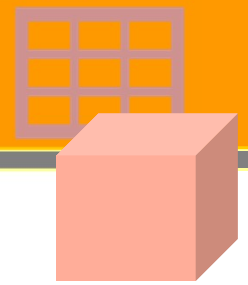


Другое представление многомерной модели данных



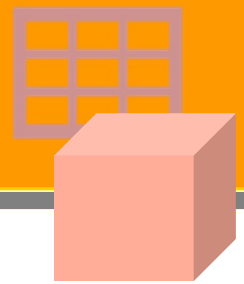


Таксономия многомерной модели данных (пример)



- **Приём абитуриентов (гиперкуб)**
  - ФАКТЫ
    - Зачислено
  - ИЗМЕРЕНИЯ
    - Календарь
      - *Год*
      - *Семестр*
    - Учебное заведение
      - *Название*
      - *Уровень*
    - География
      - *Страна*
      - *Регион*
      - *Район*

Реализация многомерных данных

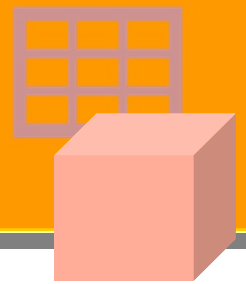


MOLAP – многомерная база данных – многомерный массив на диске с прямо адресуемыми ячейками. Агрегаты и детали могут храниться в одном гиперкубе. Как правило, урезанный атрибутный состав.

ROLAP – реляционная база данных, имитирующая гиперкуб. Агрегаты хранятся отдельно от деталей. Широкий атрибутный состав.

In memory OLAP – многомерная база данных в основной памяти, обычно в сжатом виде с хранением данных по колонкам.

Структуры данных ROLAP



В реляционном представлении многомерные данные организованы в таблицы двух видов:

1. Таблицы фактов содержат количественные данные для дальнейшего анализа:

- *транзакционные факты* на основе отдельных событий;
- *«моментальные снимки»* (snapshot) на основе состояний объекта в определённые моменты времени;
- *факты, связанные с элементами документа;*
- *факты, связанные с событиями и состоянием объекта* без подробностей.

2. Таблицы измерений предоставляют контекстную и описательную информацию о данных в таблицах фактов.

Часто представляется в виде *схемы снежинки* или *звезды*.

Структура таблицы фактов ROLAP

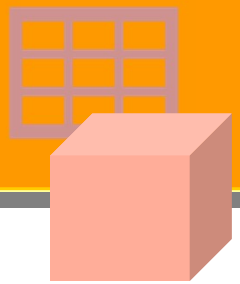


Таблица фактов содержит:

- факты одной степени детальности, соответствующие нижнему уровню иерархии
- внешние ключи таблиц измерений, которые являются составным первичным ключом.

ПРОДАЖИ

(PK, FK) Календарь_Ид
(PK, FK) Магазин_Ид
(PK, FK) Продукт_Ид
Сумма
Количество

Агрегаты вычисляются «на лету» по первичным фактам или хранятся в отдельных таблицах фактов.

Структура таблицы измерений

Таблица измерений содержит:

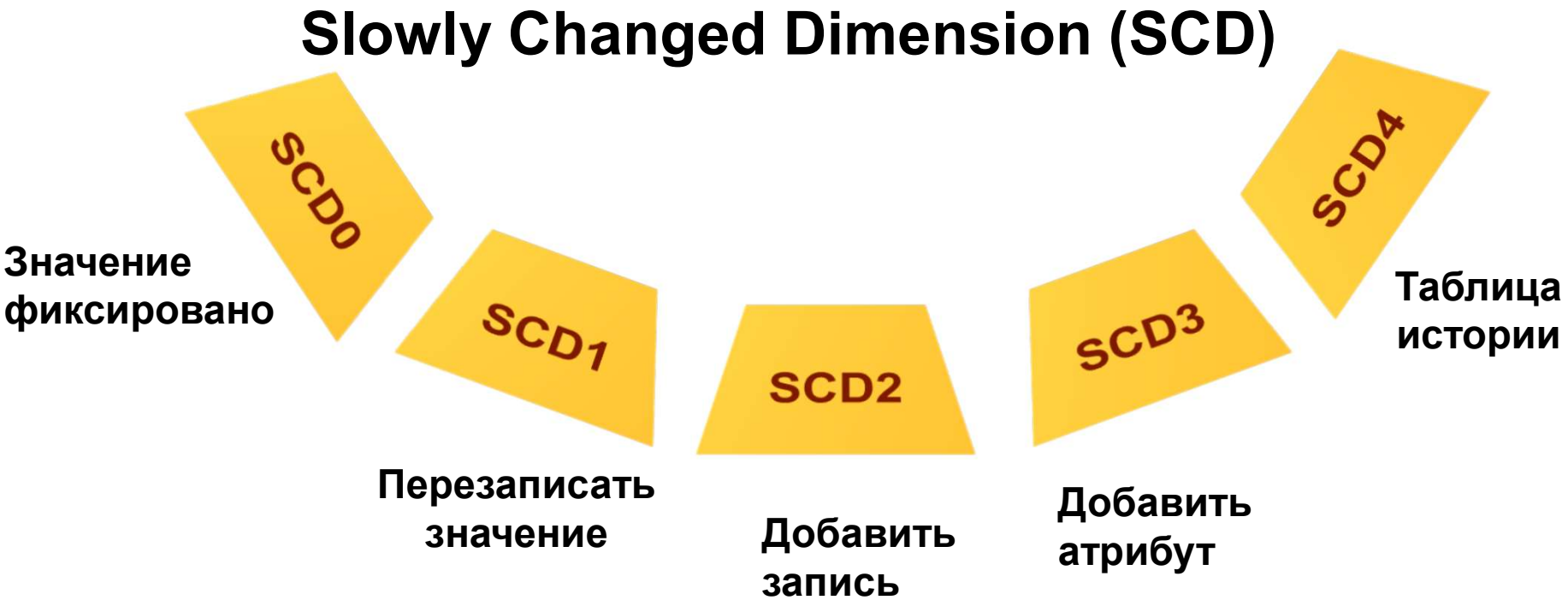
- по одной записи для каждого члена нижнего уровня
- первичный ключ записи
- характеристики нижнего уровня детальности
- квалификаторы уровней агрегирования
- все характеристики верхних уровней иерархии.

КАЛЕНДАРЬ

(PK) Месяц_Ид
Месяц_имя
Месяц_номер
Квартал_обозн
Квартал_номер
Год

В таблице измерений не рекомендуется хранить смесь разных степеней детальности, это затрудняет обработку и чревато ошибками.

Медленно меняющиеся размерности



Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)

Значение фиксировано – значение атрибута никогда не меняется. Например, *дата рождения клиента*

Перезаписать значение

Добавить запись

Добавить атрибут

Таблица истории

Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)

Значение фиксировано

Перезаписать значение – атрибуты, которые перезаписываются новыми данными. Например, *e-mail клиента или адрес доставки*

Добавить запись

Добавить атрибут

Таблица истории

Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)

Значение фиксировано

Перезаписать значение

Добавить запись –

Поддержка версий записей. Например, *категория и цена продукта*

Добавить атрибут

Таблица истории

Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)

Значение фиксировано

Перезаписать значение

Добавить запись –

Поддержка версий записей. Например, *категория и цена продукта*

Добавить атрибут

Таблица истории

Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)

Значение фиксировано

Перезаписать значение

Добавить запись

Добавить атрибут – сохраняется ограниченная история, т.е. предыдущее значение атрибута. Например, *прежнее название*

Таблица истории

Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)

Значение фиксировано

Перезаписать значение

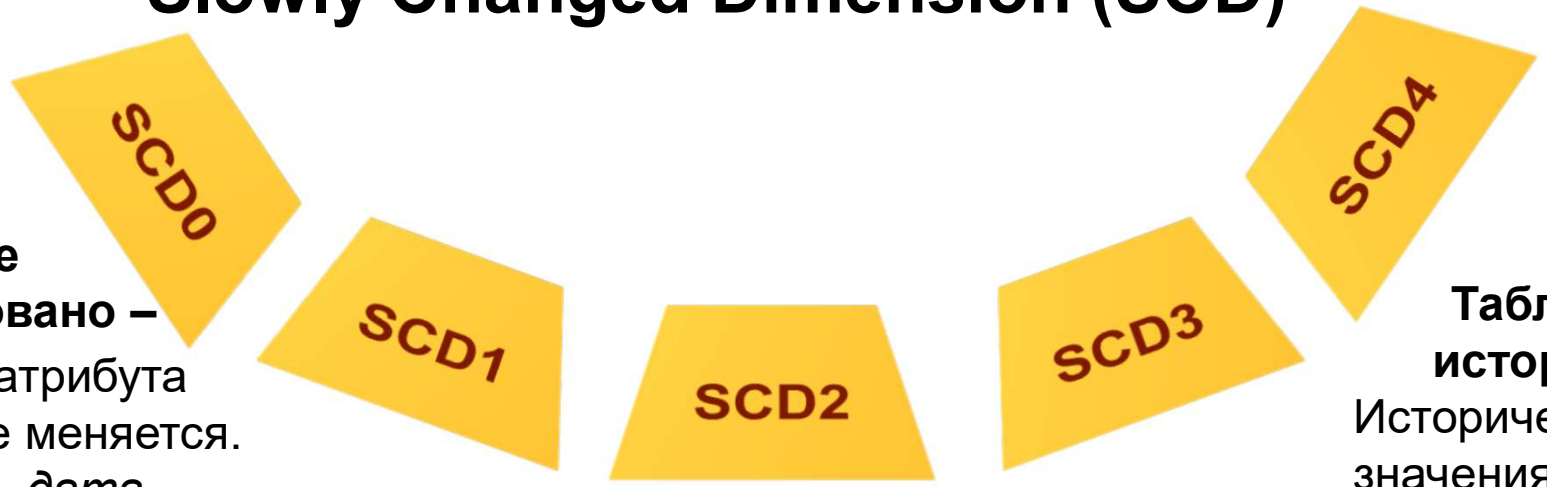
Добавить запись

Добавить атрибут

Таблица истории – Исторические значения хранятся в отдельной таблице. Например, *история продаж*.

Медленно меняющиеся размерности

Slowly Changed Dimension (SCD)



Значение фиксировано – значение атрибута никогда не меняется. Например, *дата рождения клиента*

Перезаписать значение – атрибуты, которые перезаписываются новыми данными. Например, *e-mail клиента или адрес доставки*

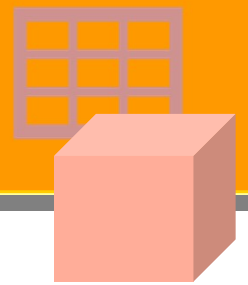
Добавить запись Поддержка версий записей Например, *категория и цена продукта*

Добавить атрибут – сохраняется ограниченная история, т.е. предыдущее значение атрибута. Например, *прежнее название*

Таблица истории – Исторические значения хранятся в отдельной таблице. Например, *история продаж*.

Может использоваться комбинация типов.

Служебные атрибуты историчности



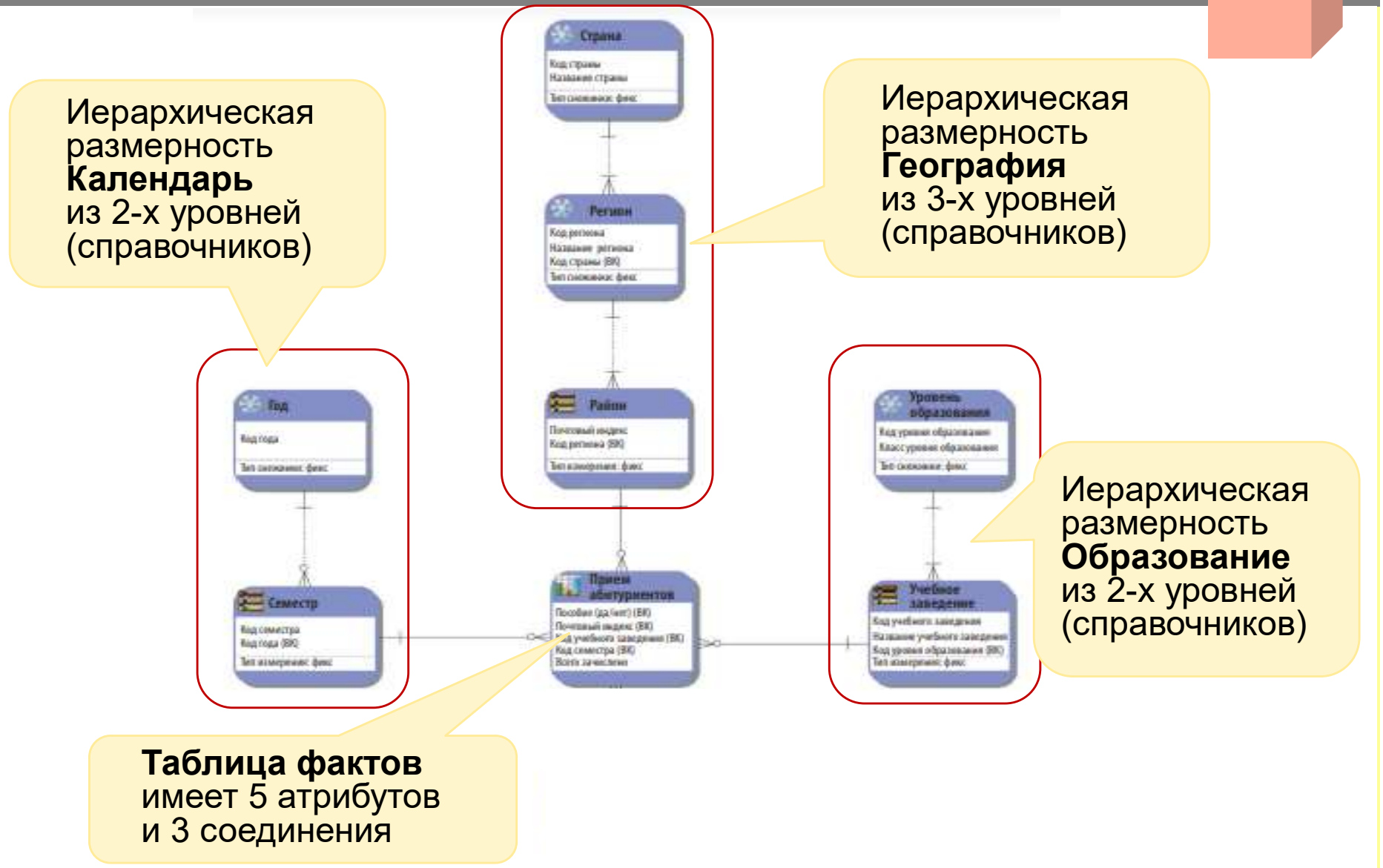
Для часто используемого способа SCD2 используются дополнительные атрибуты:

- **Effective_Date –** дата начала действия записи
- **End_Date –** дата конца действия записи
- **Actual_Flag –** состояние актуальности

Для происхождения загруженных данных используются дополнительные атрибуты:

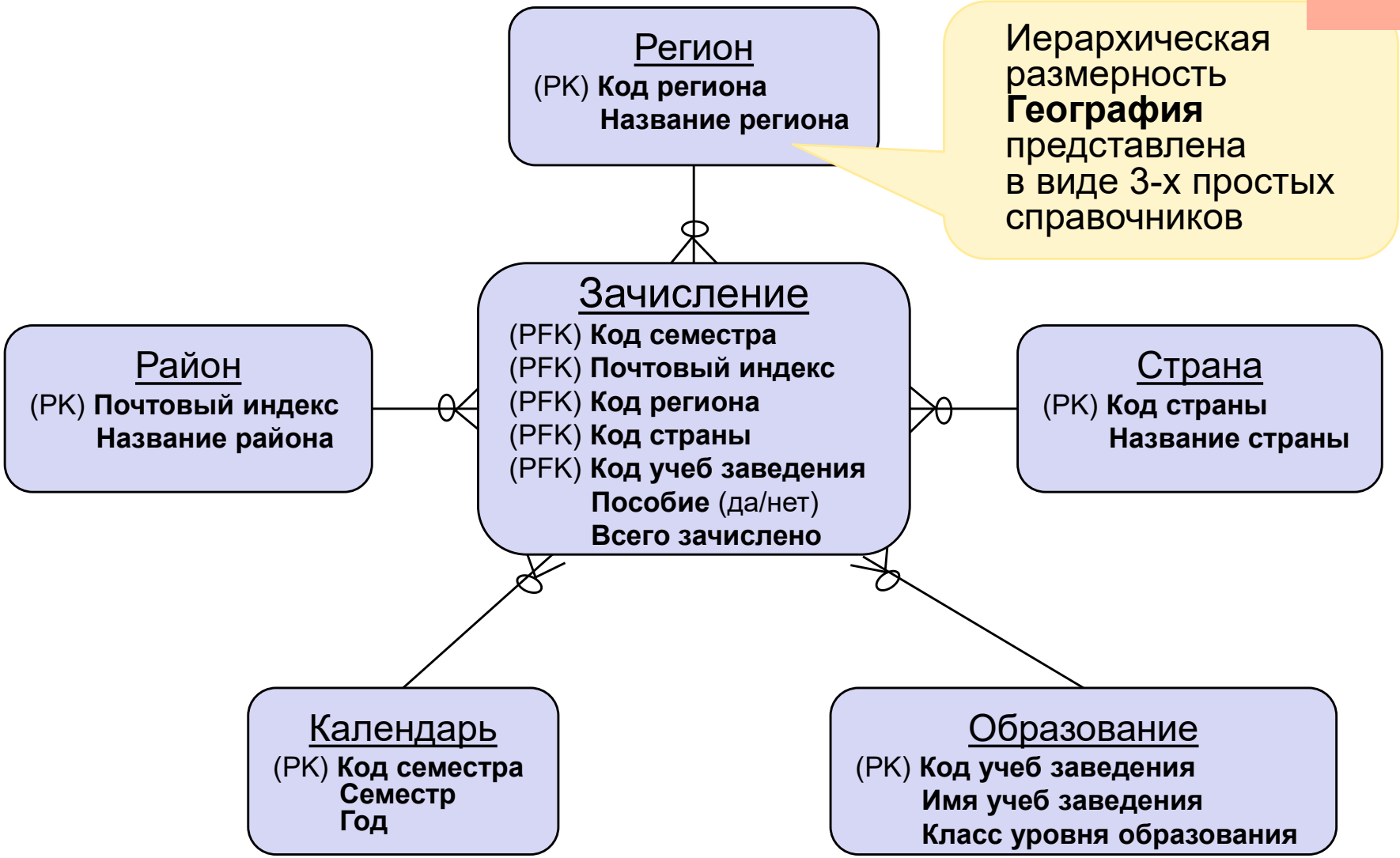
- **Load_Date_Time** – дата и время загрузки данных
- **Source_Code** – код источника данных

Схема «снежинка» (snowflake)



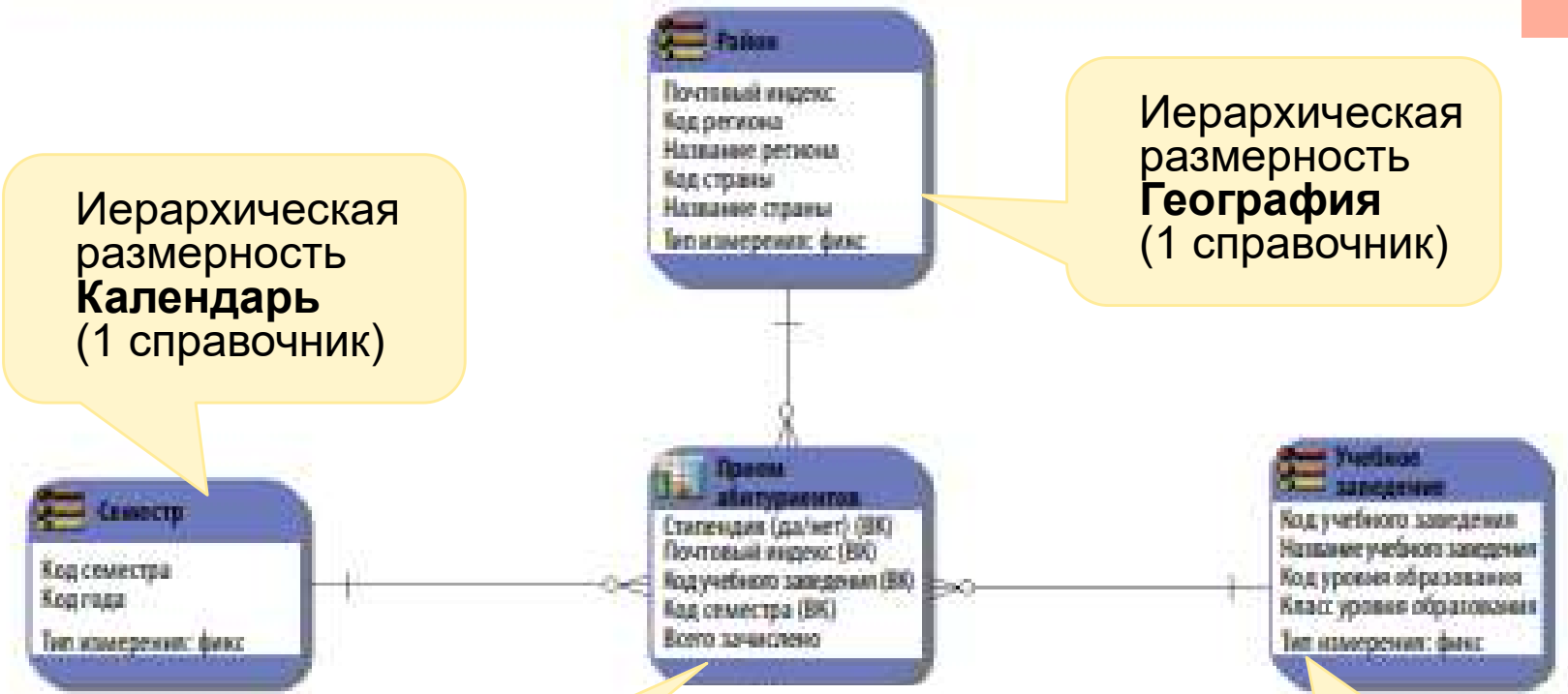
Высокая нормализация – всего 7 соединений таблиц,
3 каскадных соединения – снижение производительности

Комбинированная схема ROLAP



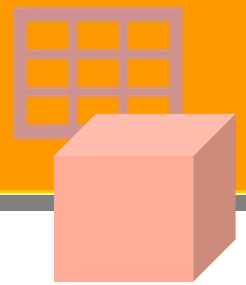
Комбинированное решение – 5 соединений без каскадов, но рост объёма хранения (7 полей вместо 5)

Схемы «звезда» (Star Schema)



Дополнительная денормализация – 3 соединения без каскадов

Максимально денормализованная схема ROLAP



Единственная плоская таблица.

Объём хранения определяется количеством атрибутов в таблице фактов

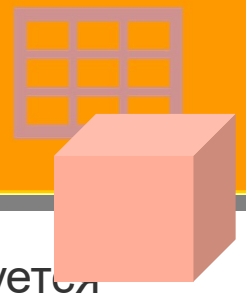
Таблица фактов имеет 11 атрибутов и не имеет соединений

Зачисление

(PK) **ID**
Семестр
Год
Почтовый индекс
Название района
Название региона
Название страны
Имя учеб заведения
Класс уровня образования
Пособие (да/нет)
Всего зачислено

Объём увеличился более чем в 2 раза (11 против 5 атрибутов), что снижает производительность, но нет операций соединения

Вместо заключения



- При анализе данных средствами бизнес-аналитики (BI/OLAP) используется *многомерное представление данных в виде гиперкубов, фактов и размерностей.*
- Бывают *простые и иерархические размерности.*
- Тематические БД для анализа организуют в виде многомерных MOLAP или чаще *реляционных витрин данных* ROLAP.
- В ROLAP витринах существуют *таблицы фактов и таблицы измерений*
- Для поддержки истории изменений имеется несколько типов *медленно меняющихся размерностей*, наиболее частый тип SCD2.
- Эти витрины создаются путём *денормализации отношений* в виде схем «снежинка», «звезда» или их комбинации, вплоть до плоской таблицы.
- Объём ROLAP-витрины определяется *размером записи таблицы фактов*, грубо числом её атрибутов (полей).
- *Время* обработки запросов существенно *зависит от количества соединений,* в большей степени *от наличия каскадов соединений,* а также *от размера записи таблицы фактов*.
- Решение находится *путём компромисса*.
- Нормализация важна для транзакционных систем, *для аналитических систем важнее денормализация отношений*.

Спасибо за внимание!

Терпения и удачи всем, кто связан с моделированием данных

Валерий Иванович Артемьев

МГТУ имени Н.Э. Баумана, каф. ИУ-5

Банк России
Департамент данных, проектов и процессов

Тел.: +7(495) 753-96-25

e-mail: avi@cbr.ru