ХРАНИЛИЩА ДАННЫХ

Храни порядок, и порядок сохранит тебя.
Латинская максима

Содержание

2

- Проблема интеграции данных
- Понятие хранилища данных
- Архитектура хранилища данных
- Реализация хранилища данных

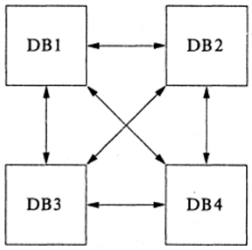
Интеграция информации

3

- *Интеграция информации* – объединение гетерогенных источников данных в единое информационное пространство, рассматриваемое пользователями как база данных.
- Методы интеграции
  - Федеративные базы данных
  - Медиаторы
  - Хранилища данных

Федеративные базы данных

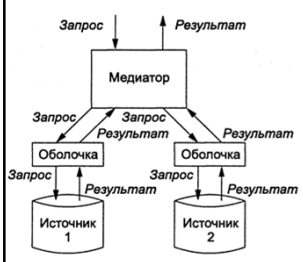
4



- *Федерация баз данных* организуется как набор API для связи каждой базы данных со всеми остальными.
- Подобная связь позволяет СУБД D_i обращаться с запросами к СУБД D_j в терминах, которые D_j воспринимает адекватно.

Медиаторы

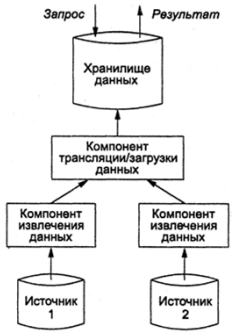
5



- *Медиатор* обеспечивает поддержку набора виртуальных таблиц, отображающих интегрированные данные из различных источников.
- Система на основе медиатора может включать в себя генератор оболочек.

Хранилища данных

6



- *Хранилище данных* предусматривает выгрузку данных из различных источников и сочетание их в рамках глобальной схемы, воспринимаемой пользователем как традиционная база данных.

Хранилище данных

7

- *Хранилище данных (Data Warehouse)* – набор данных, организованный для решения задач интеллектуального анализа данных, обладающий следующими свойствами:
  - предметная ориентированность
  - интегрированность
  - поддержка хронологии
  - неизменчивость.
- *Разделение данных*
  - базы данных – данные для оперативной обработки, источник данных для хранилища данных.
  - хранилище данных – данные для решения задач поддержки принятия решений.

Билл Инмон
р. 1945

Предметная ориентированность

8

- Организуется только для важных аспектов предметной области: *клиенты*, *товары*, *продажи* и др.
- Сфокусировано на моделировании и анализе данных *для аналитиков*, принимающих стратегические решения (не повседневные операции обработки транзакций).
- Обеспечивает *простой и краткий просмотр* предметной области путем *исключения данных, которые не являются полезными для принятия решений*.

Интегрированность

9

- Интеграция многочисленных гетерогенных источников данных
  - реляционные базы данных, txt-файлы, XML-документы и др.
- Очистка и интеграция данных
  - Обеспечение согласованности имен, семантики, единиц измерения и др. между различными источниками данных
    - Цена проживания в гостинице: валюта, налог, включение завтрака/обеда и др.
  - Преобразование данных при загрузке в хранилище.

Поддержка хронологии

10

- Временной горизонт хранилищ данных значительно больше, чем у оперативных баз данных
  - Оперативные БД: текущее значение данных
  - Хранилища данных: информация с исторической точки зрения (например, последние 5-10 лет).
- Атрибут "время"
  - Оперативные БД: может содержаться либо нет
  - Хранилище данных: всегда содержится, явно или неявно.

Неизменчивость

11

- Физически отдельное хранение данных, полученных из источников данных.
- Отсутствие операций обновления
  - Не требуются механизмы обработки транзакций, восстановления и управления параллелизмом
  - Возможные операции: *загрузка* и *чтение*.

Хранилища данных *vs* OLTP СУБД

12

- *OLTP (On-Line Transaction Processing)*
  - Основная задача традиционных РСУБД
  - Повседневные операции: покупки, склад, бухучет, платежи и др.
- *OLAP (On-Line Analytical Processing)*
  - Основная задача хранилища данных
  - Стратегические задачи: анализ данных и принятие решений
- Отличия (OLTP *vs* OLAP):
  - Ориентированность пользователей и систем: покупатель *vs* рынок
  - Содержание данных: текущие, детализированные *vs* исторические, консолидированные
  - Схема базы данных: ER *vs* "звезда"
  - Представление: текущее, локальное *vs* эволюционное, интегрированное
  - Шаблон доступа: update *vs* read-only и сложные запросы

OLTP *vs* OLAP

13

	OLTP	OLAP
Пользователи	клерки, IT-специалисты	аналитики
Функции	повседневные операции	поддержка принятия решений
БД	приложение	предмет
Данные	текущие (up-to-date) детализированные, реляционные, изолированные	исторические, агрегированные, многомерные, интегрированные
Использование	повторяющееся	*ad-hoc*
Доступ	read/write, index/hash на основе primary key	много scan
Единица работы	короткая транзакция	сложный запрос
К-во записей	10^2	10^6
К-во пользователей	10^3	10^2
Размер БД	10^2 Мб – 10^2 Гб	10^2 Гб – 10^2 Тб

Хранилища данных: почему отдельно?

14

- Настройки для высокой производительности
  - СУБД – для OLTP: методы доступа, индексирование, управление параллелизмом, восстановление
  - Хранилища – для OLAP: сложные OLAP-запросы, многомерное представление, консолидация (агрегация, суммирование)
- Разные функции, разные данные
  - *отсутствующие данные*: принятие решений требует исторических данных, которые обычно не поддерживаются источниками данных
  - *консолидация данных*: принятие решений требует консолидации (агрегации, суммирования) данных из гетерогенных источников
  - *качество данных*: различные источники данных используют несогласованные представления данных, коды и форматы

Многомерная модель данных

15

- Хранилища данных используют *многомерную модель данных*, в рамках которой данные представляются в виде куба данных.
  - *Измерение (dimension)* – набор значений атрибута
    - Поставщик={MEXX, Bvlgari, Versace, Ecco, …}
    - Товар={Одежда, Обувь, Косметика, Галантерея, …}
    - Место={Челябинск, Москва, Екатеринбург, …}
  - *Мера (measure)* – численная функция от измерений
    - Сумма: Поставщик × Товар × Место → *R*
      - Поставка(Ecco, Обувь, Челябинск)=50000 (руб.)
    - Количество: Поставщик × Товар × Место → *R*
      - Поставка(Versace, Одежда, Москва)=1 (шт.)

Проектирование хранилища данных

16

- *Таблицы измерений*
  - Измерение(ИД, Атр1, Атр2, …)
    - Поставщики(Код_П, Название, Марка, …)
    - Товары(Код_Т, Название, Цена, Скидка, …)
    - Места(Код_М, Название, Адрес, …)
- *Таблица фактов*
  - Факт(ИД_Изм1, ИД_Изм2, …, Мера1, Мера2, …)
    - Продажи(Код_П, Код_Т, Код_М, Сумма, Количество)

Проектирование хранилища данных

17

- Схемы данных
  - Звезда – таблица фактов в окружении таблиц измерений
  - Снежинка – уточнение схемы звезда, в котором выполнена нормализация таблиц измерений
  - Созвездие – множество таблиц фактов разделяют таблицы измерений

Схема "звезда"

18

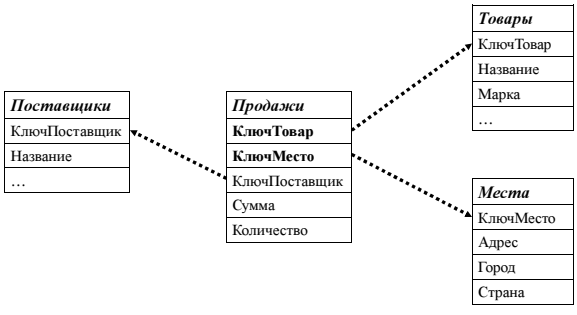


Схема "снежинка"

19

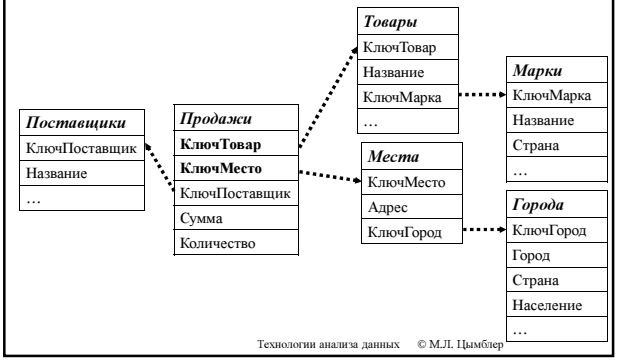
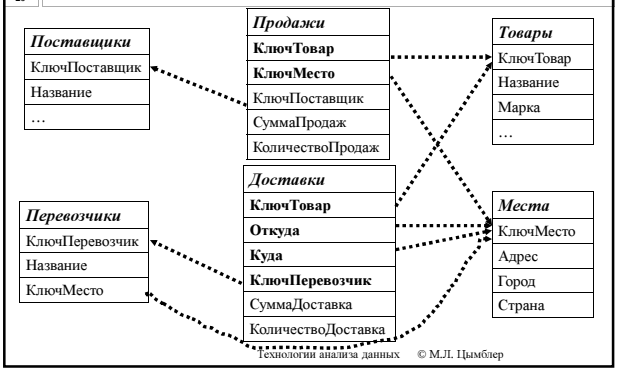


Схема "созвездие"

20



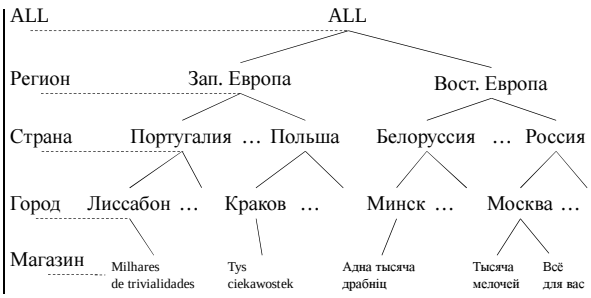
Время как измерение

21

- Дата дд.мм.гггг – это
  - номер дня в месяце
  - номер месяца в году
  - номер года
  - номер недели в году
  - номер дня в неделе
  - номер квартала
  - …
- 25.01.2010=(ИД, 25, 1, 2010, 4, 2, 1, …)

Иерархия в измерениях

22



Проектирование хранилища: 4 точки зрения

23

- Взгляд сверху-вниз
  - какая информация является релевантной и необходимой для хранилища?
- Источники данных
  - какая информация из источников данных будет помещаться в хранилище?
- Хранение данных
  - какими таблицами измерений и таблицами фактов представлено хранилище?
- Бизнес
  - какая информация требуется конечному пользователю?

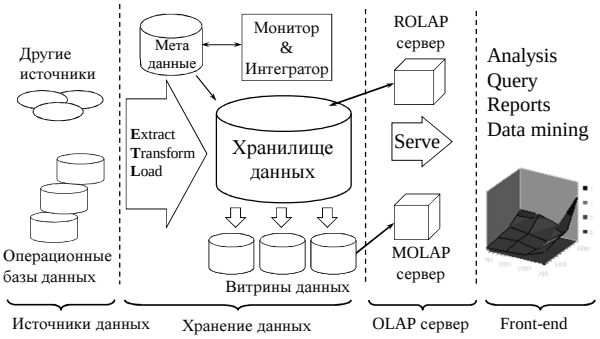
Проектирование хранилища

24

- Выбрать *бизнес-процесс* (поставки, продажи и др.)
- Выбрать *единицу (атомарный элемент данных)* бизнес-процесса
- Выбрать *измерения*
- Выбрать *меры*

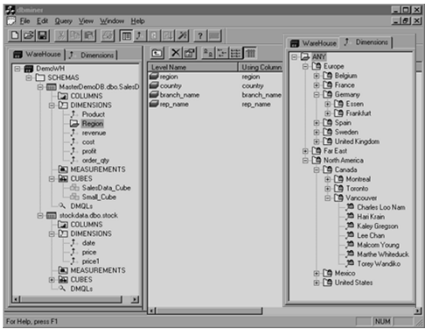
Многоуровневое хранилище данных

25



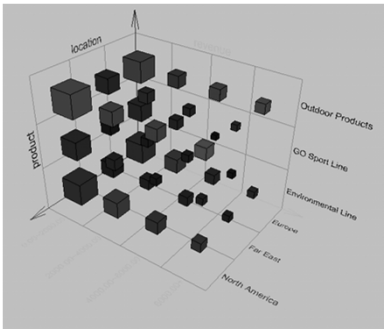
Инструментальные средства

26



Инструментальные средства

27



Язык DMQL

28

- define cube Продажи [Время, Товар, Филиал, Место]:
  - Выручка = sum(Сумма)
  - СрВыручка = avg(Сумма)
  - Вал = count(*))
- define dimension Время as (
  - КлючВремя, День, Неделя, Месяц, Квартал, Год)
- define dimension Товар as (
  - КлючТовар, НазТовар, Марка, Тип,
  - Поставщик(КлючПоставщик, ТипПоставщик))
- define dimension Филиал as (
  - КлючФилиал, НазФилиал, ТипФилиал)
- define dimension Место as (
  - КлючМесто, Улица,
  - Город(КлючГород, Округ, Страна))

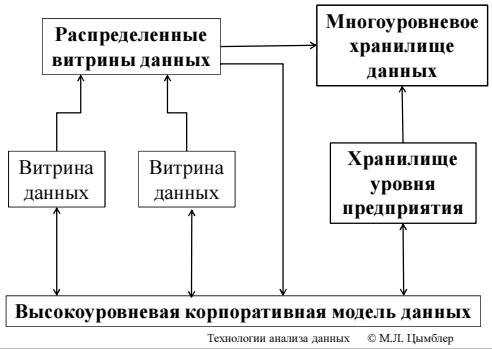
Модели хранилищ

29

- *Хранилище уровня предприятия (enterprise warehouse*)
  - Содержит данные, охватывающие все вопросы деятельности организации
- *Витрина данных (data mart)*
  - Подмножество общекорпоративных данных, которые представляют ценность для конкретных групп пользователей
- *Виртуальное хранилище (virtual warehouse)*
  - Набор представлений над операционными базами данных; лишь некоторые из этих представлений материализованы

Разработка хранилища

30



ETL

31

- *Extraction*
  - Извлечение данных из внешних гетерогенных источников
  - *Cleaning (очистка)* – определение и исправление ошибок в данных
- *Transformation*
  - Преобразование данных в формат хранилища
- *Load*
  - сортировка, суммирование, подведение итогов, создание представлений, проверка целостности, построение индексов и др.
  - *Refreshing* – распространение изменений в источниках данных на хранилище данных

Очистка: отсутствующие значения

32

- Отбрасывание записи
  - Приемлемо, если отсутствует много значений
- Ручной ввод
  - Неэффективно
- Использование глобальных констант.
  - "unknown", "n/a" – записи при кластеризации могут попасть в один класс
- Использование среднего значения или медианы
  - По всем записям
  - По записям данного класса
- Использование наиболее вероятного значения
  - Определяемое с помощью других атрибутов записи
  - Определяемое с помощью методов data mining

Очистка: "шумы" в данных

33

- *Шум* – случайная ошибка или отклонение в значениях данных. Требует *сглаживания* зашумленных значений.
- Подходы к сглаживанию
  - Binning – сортировка данных, разделение на равномощные группы, замена значений в группах на средние/медианы/границы соответствующей группы.
  - Регрессия – подбор функции, описывающей значения данных.
  - Кластеризация – определение и удаление аномалий.
  - Компьютерный метод + эксперт

Метаданные

34

- Описание структуры хранилища данных
  - Схема, представления, измерения, иерархии, определения вычисляемых данных, расположение и содержимое витрин и др.
- Операционные метаданные
  - *"Родословная данных"* (история миграции и путь трансформации), *"валюта данных"* (активные, архивные, очищенные), данные мониторинга (статистика использования хранилища, отчеты об ошибках, журналы аудита)
- Данные о производительности
  - Индексы и профили
  - Частота обновления данных
- Алгоритмы
  - Определения мер и размерности
  - Предварительно агрегированные значения и отчеты
- Бизнес-данные
  - Определения бизнес-терминов, информация о владельцах данных и административной политике

OLAP сервер

35

- *ROLAP (Relational OLAP)*
  - РСУБД или ОРСУБД, оптимизированная для хранения и обработки данных хранилища и OLAP запросов
- *MOLAP (Multidimensional OLAP)*
  - Система управления многомерными данными на основе разреженных массивов
- *HOLAP (Hybrid OLAP)*
  - Гибрид: низкий уровень – реляционные данные, высокий уровень – массивы

Использование хранилищ данных

36

- OLTP
  - обычные запросы, статистический анализ и отчеты на основе кросс-таблиц, диаграмм и графиков
- OLAP
  - анализ многомерных данных
  - основные OLAP операции
- Data mining
  - определение скрытых закономерностей
  - ассоциативные правила, классификация, кластеризация, …, визуализация

Заключение

37

- Интеграция информации – объединение гетерогенных источников данных в единое информационное пространство.
- Методы интеграции
  - Федеративные базы данных
  - Медиаторы
  - Хранилища данных
- Хранилище данных – набор данных, организованный для задач поддержки принятия решений, обладающий следующими свойствами:
  - предметная ориентированность
  - интегрированность
  - поддержка хронологии
  - неизменчивость.