ХРАНИЛИЩА ДАННЫХ И OLAP

Храни порядок, и порядок сохранит тебя.
Латинская максима

Технологии анализа данных

Содержание

2

- Подходы к интеграции информации
  - Федеративная база данных
  - Медиатор
  - Хранилище данных
- Построение хранилища данных
- Оперативный анализ данных (OLAP)

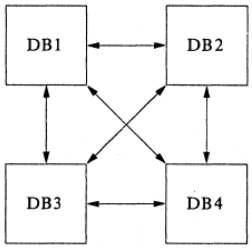
Интеграция информации

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

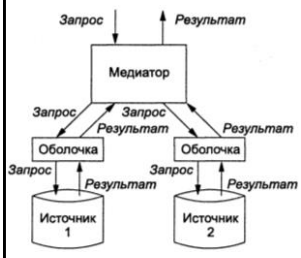
4



- *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

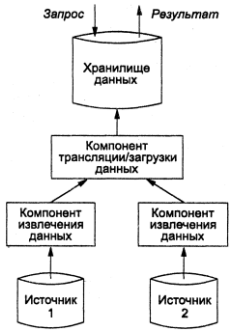
5



- *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*.

Использование хранилищ данных

12

- Оперативная обработка транзакций (OLTP, Online Transaction Processing)
  - обычные запросы на обновление данных, статистические отчеты
- Оперативный анализ данных (OLAP, Online Analytical Processing)
  - сложные запросы на выборку по многим критериям с использованием функций агрегации
- Интеллектуальный анализ данных (Data mining)
  - определение скрытых закономерностей в данных

Хранилища данных *vs* OLTP СУБД

13

- *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

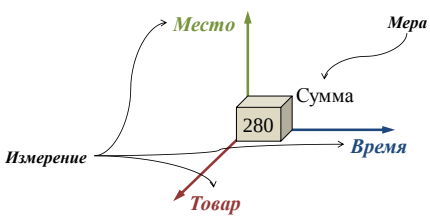
14

	OLTP	OLAP
Пользователи	клерки, IT-специалисты	аналитики
Функции	повседневные операции	поддержка принятия решений
Данные	текущие, детализированные, реляционные	исторические, агрегированные, многомерные
Использование	повторяющееся	незапланированное
Доступ	read/write	scan
Единица работы	короткая транзакция	сложный запрос на выборку
Порядок кол-ва записей	10^2	10^6
Порядок кол-ва пользователей	10^3	10^2
Порядок размера данных	10^2 Мб – 10^2 Гб	10^2 Гб – 10^2 Тб

Многомерная модель данных

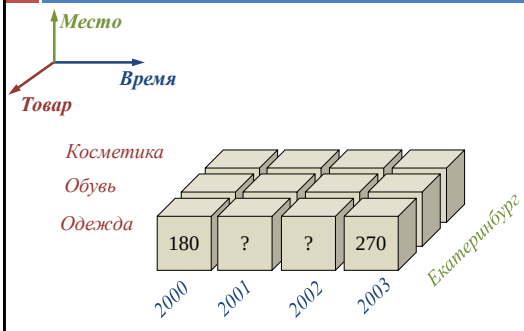
15

- Данное рассматривается как факт с численными параметрами и текстовыми измерениями, характеризующими данный факт.



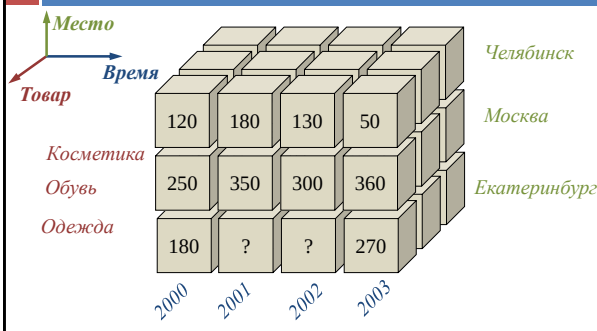
Куб данных

16



Куб данных

17



Многомерная модель данных

18

- 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

19

- *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

20

- Схемы данных
  - *Звезда (star)* – таблица фактов в окружении таблиц измерений
  - *Снежинка (snowflake)* – уточнение схемы звезда, в котором выполнена нормализация таблиц измерений
  - *Созвездие (constellation)* – множество таблиц фактов разделяют таблицы измерений

Схема "звезда"

21

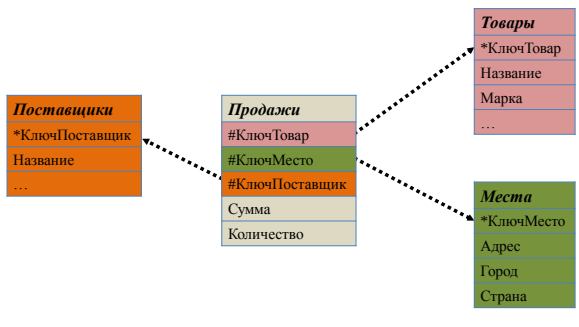


Схема "снежинка"

22

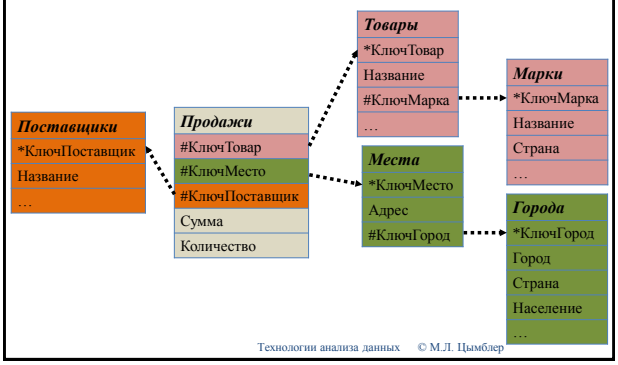
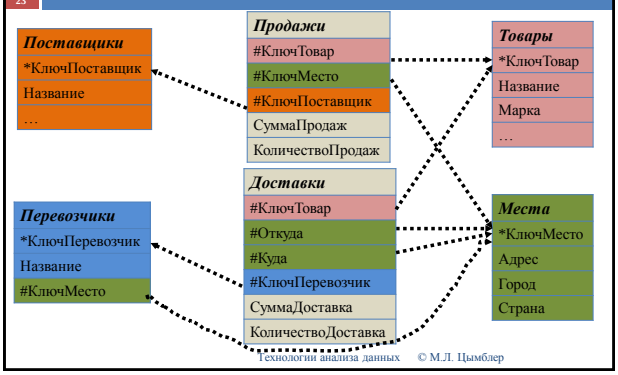


Схема "созвездие"

23



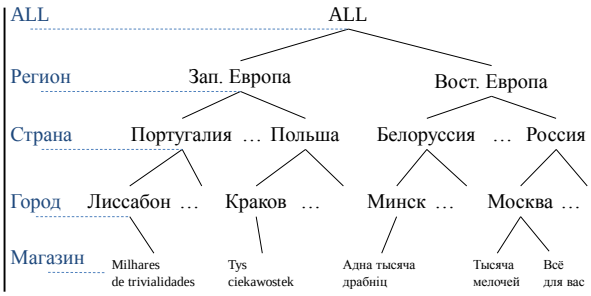
Время как измерение

24

- 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

25



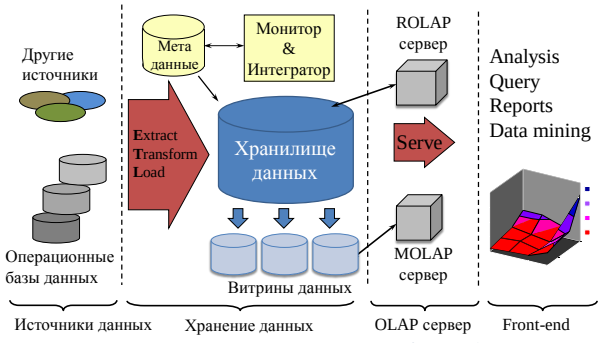
Проектирование хранилища

26

- Отбор релевантных данных
  - 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?

Многоуровневое хранилище данных

27



ETL

28

- *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

Метаданные

29

- 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

30

- *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

Язык DMQL

31

- 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))

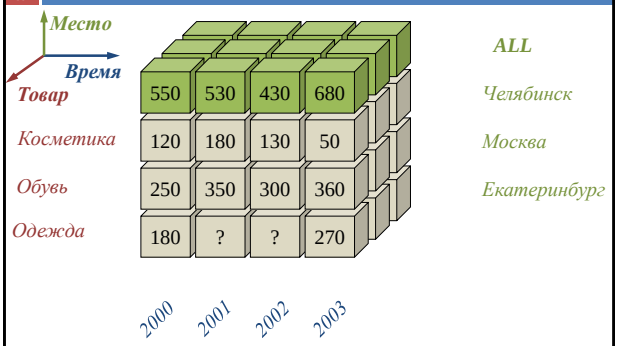
OLAP-куб

32

- *OLAP-куб* представляет собой куб данных, в котором каждое измерение дополняется значением *ALL* и полученные таким образом новые точки пространства вычисляются с помощью заданной *агрегатной функции*.
- Агрегатные функции
  - *Дистрибутивные*
    - count(), sum(), min(), max() и др.
  - *Алгебраические*
    - avg(), stddev() и др.
  - *Холистические меры*
    - median(), mode() и др.

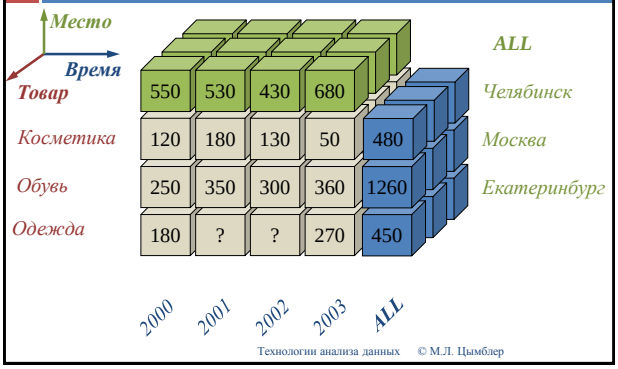
OLAP-куб

33



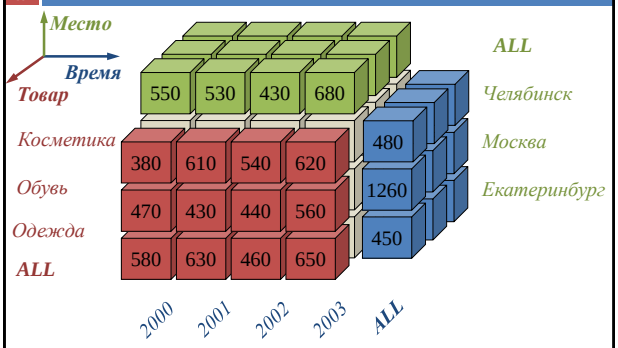
OLAP-куб

34



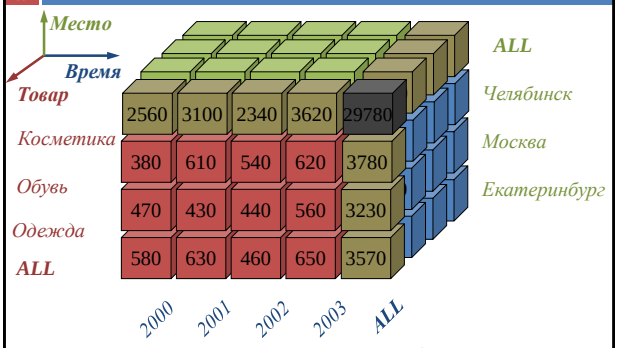
OLAP-куб

35



OLAP-куб

36



OLAP-куб и SQL

37

- ROLLUP BY
  - вычисление агрегата меры для каждого указанного измерения
  - вычисление частичных итогов (справа налево в списке группируемых измерений)
  - вычисление общего итога
- CUBE BY
  - вычисление агрегата меры для всех возможных комбинаций указанных измерений

ROLLUP BY

38

```
select Время, Место, Товар,
  sum(Сумма) as Прибыль
from Продажи
rollup by (Время, Место, Товар)
```

```
select Время, Место, Товар,
  sum(Сумма) as Прибыль
from Продажи
group by (Время, Место, Товар)
union
select Время, Место, '',
  sum(Сумма) as Прибыль
from Продажи
group by (Время, Место)
union
select Время, '', '',
  sum(Сумма) as Прибыль
from Продажи
group by (Время)
union
select '', '', '', sum(Сумма) as Прибыль
from Продажи
```


ROLLUP BY

39

Время	Место	Товар	Сумма
2000	Челябинск	Одежда	100
2000	Челябинск	Косметика	120
2000	Москва	Одежда	250
2000	Москва	Косметика	75
2001	Челябинск	Одежда	230
2001	Челябинск	Косметика	310
2001	Москва	Одежда	170
2001	Москва	Косметика	350

ROLLUP BY

40

```
select
   Время, Место, Товар,
   sum(Сумма) as Прибыль
from Продажи
rollup by (Время,
          Место, Товар)
```

Время	Место	Товар	Прибыль
2000	**Челябинск**	**Одежда**	100
2000	Челябинск	**Косметика**	120
2000	**Челябинск**	[NULL]	**220**
2000	**Москва**	**Одежда**	250
2000	Москва	**Косметика**	75
2000	**Москва**	[NULL]	**325**
2000	[NULL]	[NULL]	**545**
2001	**Челябинск**	**Одежда**	230
2001	Челябинск	**Косметика**	310
2001	**Челябинск**	[NULL]	**540**
2001	**Москва**	**Одежда**	170
2001	Москва	**Косметика**	350
2001	**Москва**	[NULL]	**520**
2001	**[NULL]**	**[NULL]**	**1 060**
[NULL]	**[NULL]**	**[NULL]**	**1 605**

CUBE BY

41

Время	Место	Товар	Сумма
2000	Челябинск	Одежда	100
2000	Челябинск	Косметика	120
2000	Москва	Одежда	250
2000	Москва	Косметика	75
2001	Челябинск	Одежда	230
2001	Челябинск	Косметика	310
2001	Москва	Одежда	170
2001	Москва	Косметика	350

CUBE BY

42

```
select
   Время, Место, Товар,
   sum(Сумма) as Прибыль
from Продажи
cube by (Время,
          Место, Товар)
```

Время	Место	Товар	Прибыль
2000	**Челябинск**	**Одежда**	100
2000	Челябинск	**Косметика**	120
2000	**Челябинск**	**[NULL]**	**220**
2000	**Москва**	**Одежда**	250
2000	Москва	**Косметика**	75
2000	**Москва**	**[NULL]**	**325**
2000	[NULL]	**Одежда**	**350**
2000	[NULL]	**Косметика**	**195**
2000	**[NULL]**	**[NULL]**	**545**
2001	**Челябинск**	**Одежда**	230
2001	Челябинск	**Косметика**	310
2001	**Челябинск**	**[NULL]**	**540**
2001	**Москва**	**Одежда**	170
2001	Москва	**Косметика**	350
2001	**Москва**	**[NULL]**	**520**

CUBE BY

43

```
select
  Время, Место, Товар,
  sum(Сумма) as Прибыль
from Продажи
cube by (Время,
         Место, Товар)
```

Время	Место	Товар	Прибыль
[NULL]	Челябинск	Одежда	330
[NULL]	Челябинск	Косметика	430
[NULL]	Челябинск	[NULL]	760
[NULL]	Москва	Одежда	420
[NULL]	Москва	Косметика	425
[NULL]	Москва	[NULL]	845
[NULL]	[NULL]	Одежда	750
[NULL]	[NULL]	Косметика	855
[NULL]	[NULL]	[NULL]	1 605

GROUPING

44

- Позволяет отличить NULL-значения, добавляемые ROLLUP/CUBE, от «настоящих» NULL-значений.

```
SELECT dim1, dim2,
    SUM(measure) as val,
    GROUPING(dim1) as f1_isALL,
    GROUPING(dim2) as f2_isALL
FROM FactTab
GROUP BY CUBE (dim1, dim2)
ORDER BY dim1, dim2;
```

DIM1	DIM2	VAL	F1_isALL	F2_isALL
1	1	4363.55	0	0
1	2	4794.76	0	0
1	3	4718.25	0	0
1	4	5387.45	0	0
1	5	5027.34	0	0
1		**24291.35**	0	**1**
2	1	5652.84	0	0
2	2	4583.02	0	0
2	3	5555.77	0	0
2	4	5936.67	0	0
2	5	4508.74	0	0
2		**26237.04**	0	**1**
	1	**10016.39**	**1**	0
	2	**9377.78**	**1**	0
	3	**10274.02**	**1**	0
	4	**11324.12**	**1**	0
	5	**9536.08**	**1**	0
		50528.39	**1**	**1**

GROUPING

45

```
SELECT dim1, dim2,
    SUM(measure) AS val,
    GROUPING(dim2) AS f1_isALL,
    GROUPING(dim2) AS f2_isALL
FROM   FactTab
GROUP BY CUBE (dim1, dim2)
HAVING GROUPING(dim1) = 1 OR
GROUPING(dim2) = 1
ORDER BY GROUPING(dim1),
GROUPING(dim2);
```

DIM1	DIM2	VAL	F1_isALL	F2_isALL
1		24291.35	0	1
2		26237.04	0	1
	1	10016.39	1	0
	2	9377.78	1	0
	3	10274.02	1	0
	4	11324.12	1	0
	5	9536.08	1	0
		50528.39	1	1

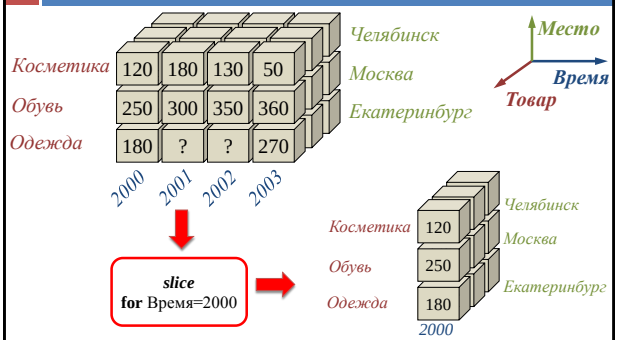
OLAP-операции (построение нового OLAP-куба)

46

- *Срез (slice and dice)*
  - проекция и/или отбор
- *Агрегация (roll-up, drill-up)*
  - вычисление меры при продвижении измерения снизу вверх по иерархии
- *Детализация (drill-down, roll-down)*
  - вычисление меры при продвижении измерения сверху вниз по иерархии
- *Вращение (pivot)*
  - изменение порядка представления (визуализации) измерений.

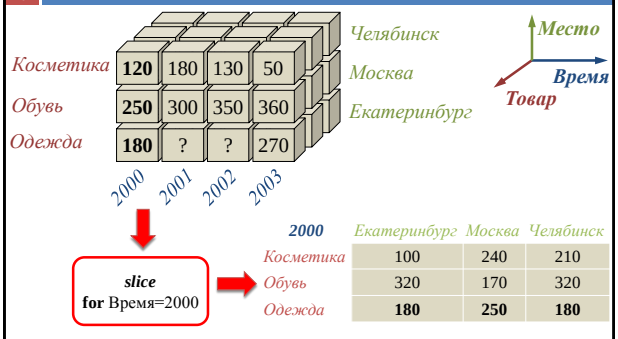
Срез (slice)

47



Срез (slice)

48

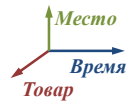


2000	Екатеринбург	Москва	Челябинск
Косметика	100	240	210
Обувь	320	170	320
Одежда	**180**	**250**	**180**

Вращение (pivot)

49

2000	Екатеринбург	Москва	Челябинск
Косметика	100	240	210
Обувь	320	170	320
Одежда	180	250	120

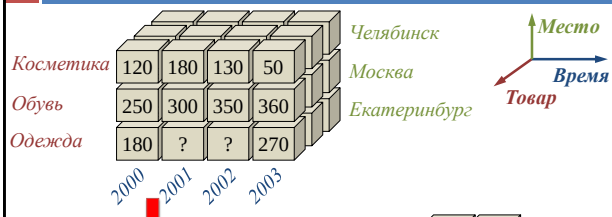


pivot

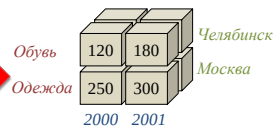
2000	Одежда	Обувь	Косметика
Екатеринбург	180	320	100
Москва	250	170	240
Челябинск	120	320	210

Срез (dice)

50

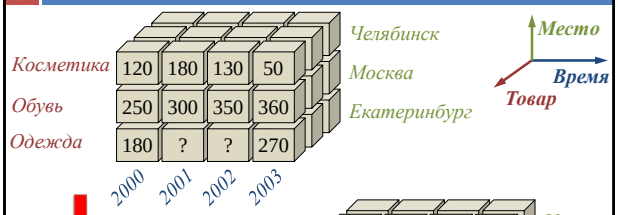


dice
for (Время=2000 or Время=2001) and (Место=Челябинск or Место=Москва) and (Товар=Обувь or Товар=Одежда)

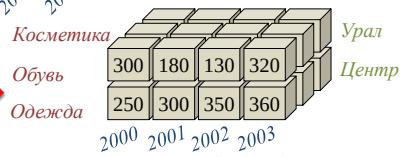


Агрегация (roll-up)

51

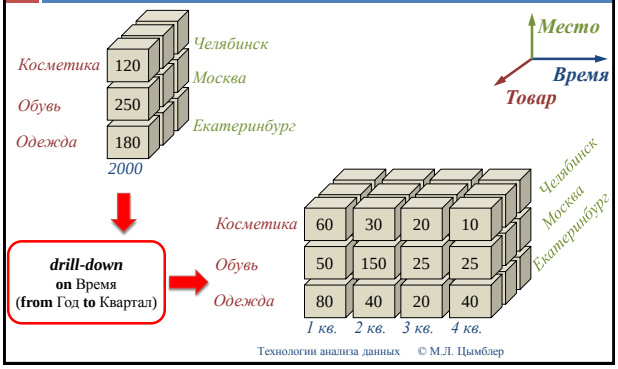


roll-up
on Место
(**from** Город **to** Округ)



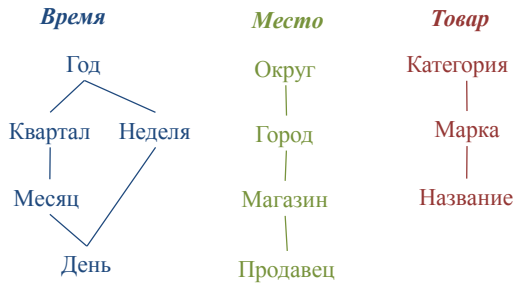
Детализация (drill-down)

52



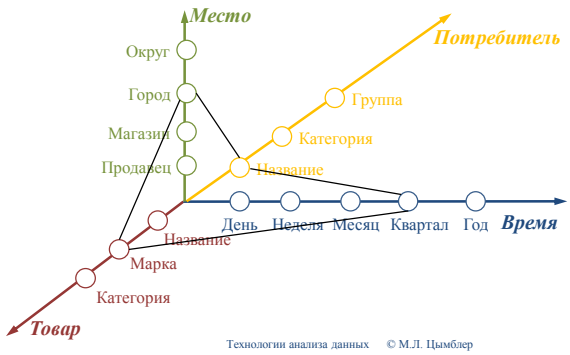
Иерархия в измерениях

53



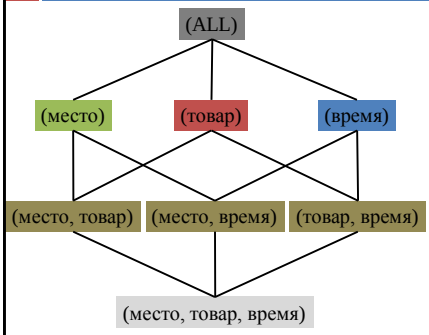
Модель OLAP-запросов

54



Решетки кубоидов

55



0-D (apex)
group by ()

1-D
group by (место)
group by (товар)
group by (время)

2-D
group by (место, товар)
group by (место, время)
group by (товар, время)

3-D (base)
group by
(место, товар, время)

ЧаВо о кубах

56

- Объем куба

$$V_{куб} = \prod_{i=1}^{n} d_i$$

- Объем OLAP-куба

$$V_{OLAP-куб} = \prod_{i=1}^{n} (m + d_i)$$

- Количество кубоидов в OLAP-кубе

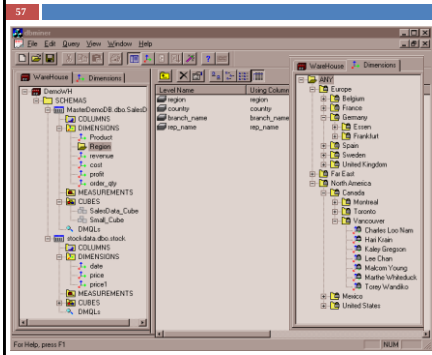
$$Q_1 = 2^n \qquad Q_L = \prod_{i=1}^{n} (1 + L_i)$$

- Как визуализировать k-мерный куб (k>3)?
  Как набор из d_k (k-1)-мерных кубов.

- n – размерность (количество измерений)
- d_i – количество значений в i-м измерении
- m – количество мер
- Li – количество уровней иерархии в i-м измерении

Инструментальные средства

57



Инструментальные средства

58