Поддержка хранилищ данных и анализ данных в СУБД Oracle

Не счесть алмазов в каменных пещерах...

А. Пушкин

Содержание

- Понятия хранилища данных, анализа данных
- Поддержка хранилища данных и анализа данных в СУБД Oracle

Хранилища данных

- *Хранилище данных (Data Warehouse)* содержит неизменчивую интегрированную предметно-ориентированную совокупность *исторических данных* крупной компании с целью поддержки принятия *стратегических решений*.
- Основные идеи:
  - интеграция и согласование ранее разъединенных детализированных данных;
  - разделение данных для операционной обработки и данных для решения задач анализа.
- Основные замечания:
  - хранилище данных – это не анализ данных, а подготовка данных к анализу;
  - хранилище данных определяет не архитектуру построения аналитической системы, представление данных, СУБД и т.п., а процессы по сбору данных в эту систему;
  - хранилище данных предполагает реализацию единого централизованного источника данных компании.

Витрины данных

- *Витрина данных (Data Mart)* – специализированное хранилище данных, связанное с какими-либо *конкретными аспектами деятельности* компании.
- Используется для поддержки принятия решений в соответствующей сфере деятельности компании.
- Создается на основе общего хранилища либо независимо.

Анализ данных

Разновидность / Характеристика	**Статический анализ (DSS, Decision Support Systems)**	**Динамический анализ (OLAP, On-Line Analytical Processing)**
Назначение	Регламентированная аналитическая обработка	Моделирование и построение прогнозов
Запросы к данным	Предсказуемые – "сколько?", "когда?"	Непредсказуемые (*ad hoc*) – "почему?", "что будет/было, если?"
Время отклика	Не регламентируется	Секунды
Типичные операции	Регламентированный отчет, диаграмма	Последовательность отчетов, диаграмм, экранных форм. Динамическое изменение уровней агрегации и срезов данных.
Уровень агрегации данных	Детализированные и суммарные	Как правило, суммарные

Реализация хранилищ данных: основные проблемы

- Гетерогенная аппаратно-программная среда
- Распределенность данных
- Организация и поддержка единого словаря данных
- Защита данных
- Эффективное хранение и обработка очень больших объемов данных
  - Классификация хранилищ по объему данных:
    малые – до 3 Гб и 10^9 записей в таблице;
    средние – до 25 Гб и 10^{11} записей;
    большие – до 200 Гб и 10^{12} записей;
    очень большие – свыше 200 Гб и 10^{12} записей.
  - Выбор модели данных и СУБД: реляционная или *многомерная*?

Многомерное представление данных

Реляционное представление

Время	Филиал	Прибыль
2000	Екатеринбург	280 000
2000	Москва	250 000
2000	Челябинск	120 000
2001	Екатеринбург	160 000
2001	Москва	350 000
2001	Челябинск	230 000
2002	Екатеринбург	310 000
2002	Челябинск	130 000
2003	Екатеринбург	270 000
2003	Москва	140 000

2-мерные представления (2-мерные гиперкубы)

Филиал	2000	2001	2002	2003
Екатеринбург	280 000	160 000	310 000	270 000
Москва	250 000	350 000	–	140 000
Челябинск	120 000	230 000	130 000	–

Время	Екатеринбург	Москва	Челябинск
2000	280 000	250 000	120 000
2001	160 000	350 000	230 000
2002	310 000	–	130 000
2003	270 000	140 000	–

Корпоративные системы баз данных © М.Л. Цымблер 7

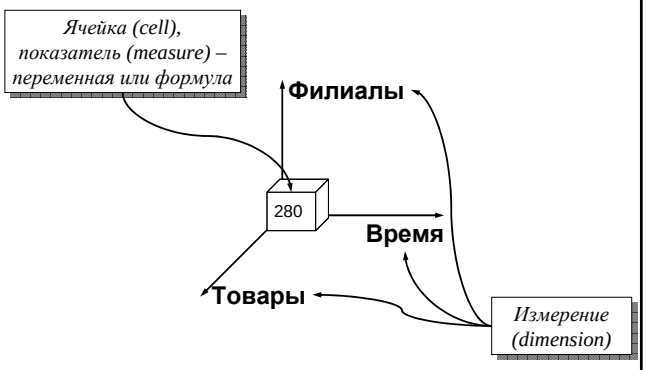
Многомерное представление данных

Время	Филиал	Товар	Прибыль
2000	Екатеринбург	Косметика	280 000
2000	Екатеринбург	Обувь	330 000
2000	Екатеринбург	Одежда	180 000
2000	Москва	Обувь	220 000
2000	Москва	Одежда	250 000
2000	Челябинск	Одежда	120 000
2001	Екатеринбург	Косметика	70 000
2001	Екатеринбург	Обувь	160 000
2001	Москва	Косметика	40 000
2001	Москва	Одежда	350 000
2001	Челябинск	Обувь	230 000
2002	Екатеринбург	Косметика	310 000
2002	Челябинск	Одежда	130 000
2003	Екатеринбург	Одежда	270 000
2003	Москва	Обувь	140 000
2003	Москва	Одежда	360 000

Реляционное представление

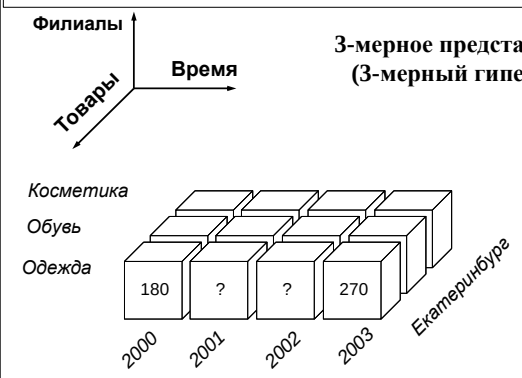
Корпоративные системы баз данных © М.Л. Цымблер 8

Многомерное представление данных



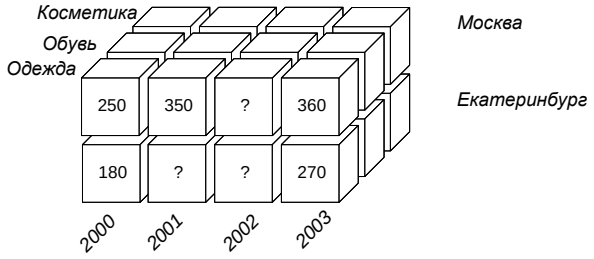
Корпоративные системы баз данных © М.Л. Цымблер 9

Многомерное представление данных



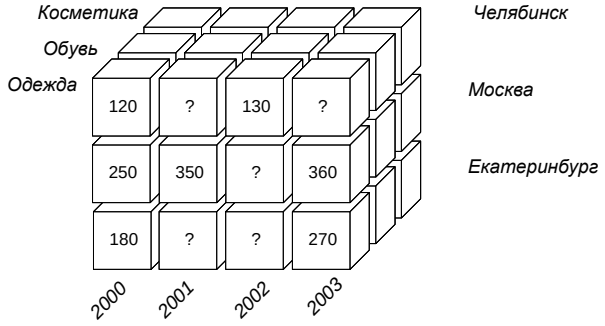
Корпоративные системы баз данных © М.Л. Цымблер 10

Многомерное представление данных



Корпоративные системы баз данных © М.Л. Цымблер 11

Многомерное представление данных



Корпоративные системы баз данных © М.Л. Цымблер 12

Поликубическая vs. гиперкубическая модели

- *Поликубическая модель* предполагает возможность определения в многомерной базе данных нескольких гиперкубов с различной размерностью и показателями.
- *Гиперкубическая модель* предполагает, что в многомерной базе данных может существовать только один гиперкуб и все показатели должны определяться одним и тем же набором измерений.

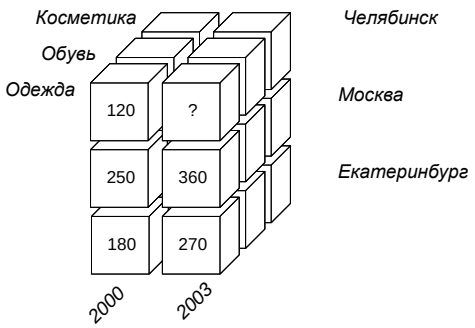
Поликубическая vs. гиперкубическая модели

- Пример многомерной базы данных:
  - Измерения
    - Время: День → Месяц → Год
    - Филиал: Продавец → Город
    - Товар: Тип → Производитель → Страна
  - Показатели:
    - Прибыль
    - Рабочее время продавца
- Поликубическая модель:
  - 2-мерный куб {День, Продавец} ⇒ Рабочее время продавца
    3-мерный куб {День, Продавец, Товар} ⇒ Прибыль
- Гиперкубическая модель:
  - 3-мерный куб {День, Продавец, Товар} ⇒
    Прибыль, Рабочее время продавца

Операции манипулирования многомерными данными

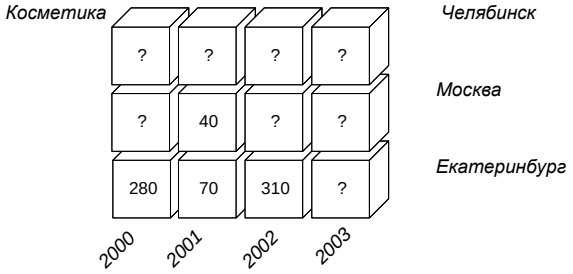
- *Срез (slice)* – построение из куба данных нового куба путем фиксации значения одного или более измерений.
- *Вращение (rotation)* – изменение порядка представления (визуализации) измерений.
- *Агрегация (drill-up)* – построение из куба данных (или среза) нового куба (среза) путем агрегирования показателей по каким-либо измерениям, где на множестве измерений определены *отношения иерархии*.
- *Детализация (drill-down)* – операция, обратная к операции агрегации.

Срез



Корпоративные системы баз данных © М.Л. Цымблер 16

Срез



Корпоративные системы баз данных © М.Л. Цымблер 17

Вращение

Филиал	2000	2001	2002	2003
Екатеринбург	280 000	160 000	310 000	270 000
Москва	250 000	350 000	–	140 000
Челябинск	120 000	230 000	13	

Филиал	2003	2002	2001	2000
Екатеринбург	270 000	310 000	160 000	280 000
Москва	140 000	–	350 000	250 000
Челябинск	–	130 000	230 000	120 000

Время	Екатеринбург	Москва	Челябинск
2000	280 000	250 000	120 000
2001	160 000	350 000	230 000
2002	310 000	–	130 000
2003	270 000	140 000	

Время	Челябинск	Екатеринбург	Москва
2000	120 000	280 000	250 000
2001	230 000	160 000	350 000
2002	130 000	310 000	–
2003	–	270 000	140 000

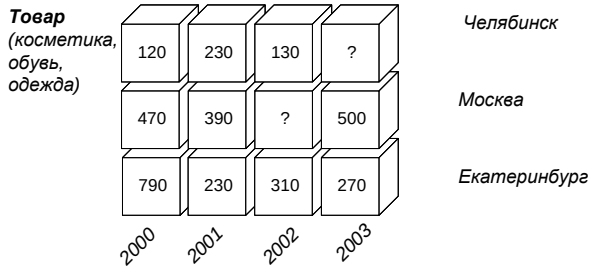
Корпоративные системы баз данных © М.Л. Цымблер 18

Иерархия в измерениях

- Примеры отношений иерархии
  - Время:
    День → Неделя → Месяц → Квартал → Год
  - Филиал:
    Продавец → Магазин → Город → Регион
  - Товар:
    Тип → Производитель → Страна
- Многоуровневая иерархия порождает гиперкуб большого объема.
  Например (для небольшой фирмы):
  365 дней * 5 лет *
  10 продавцов * 4 магазина * 3 города * 2 региона *
  3 типа * 20 производителей * 3 страны ≈ **80 млн. ячеек**
- Полученный гиперкуб может содержать "пустоты".
  Например:
  65 выходных и праздничных дней * 5 лет *
  10 продавцов * 4 магазина * 3 города * 2 региона *
  3 типа * 20 производителей * 3 страны ≈ **14 млн. ячеек (17,5%)**

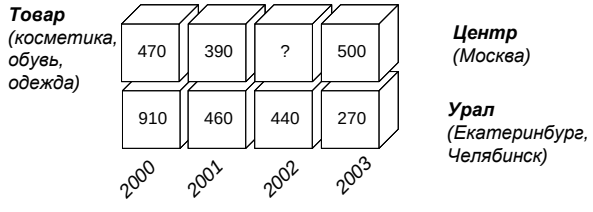
Корпоративные системы баз данных © М.Л. Цымблер 19

Агрегация



Корпоративные системы баз данных © М.Л. Цымблер 20

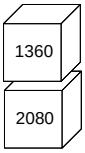
Агрегация



Корпоративные системы баз данных © М.Л. Цымблер 21

Агрегация

Товар *(косметика, обувь, одежда)*



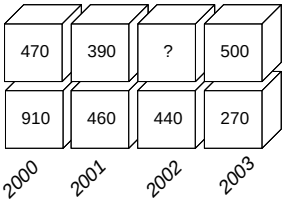
Центр *(Москва)*

Урал *(Екатеринбург, Челябинск)*

Время *(2000, 2001, 2002, 2003)*

Детализация

Товар *(косметика, обувь, одежда)*

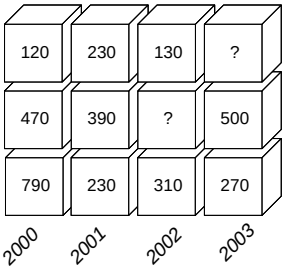


Центр *(Москва)*

Урал *(Екатеринбург, Челябинск)*

Детализация

Товар *(косметика, обувь, одежда)*

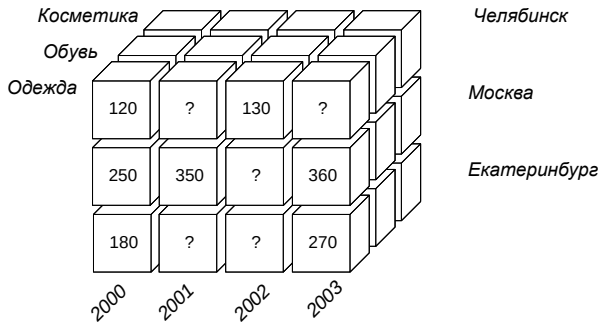


Челябинск

Москва

Екатеринбург

Детализация



Корпоративные системы баз данных © М.Л. Цымблер 25

Средства анализа данных в Oracle

- Фраза ROLLUP – вычисление промежуточных и окончательных итогов по указанным полям таблицы.
- Фраза CUBE – вычисление всех возможных промежуточных и окончательных итогов по указанным полям таблицы.
- Запросы TOP-N – список из N записей, имеющих в указанном поле таблицы максимальное/минимальное значение.

Корпоративные системы баз данных © М.Л. Цымблер 26

Модельное хранилище данных

Время	Филиал	Товар	Прибыль
2000	Челябинск	Одежда	100 000
2000	Челябинск	Косметика	120 000
2000	Москва	Одежда	250 000
2000	Москва	Косметика	75 000
2001	Челябинск	Одежда	230 000
2001	Челябинск	Косметика	310 000
2001	Москва	Одежда	170 000
2001	Москва	Косметика	350 000

Корпоративные системы баз данных © М.Л. Цымблер 27

Анализ данных с помощью ROLLUP

```
select
   Время, Филиал,
   Товар, sum(Прибыль)
   as Прибыль
from Продажи
group by
    rollup (Время,
          Филиал, Товар)
```

Время	Филиал	Товар	Прибыль
2000	**Челябинск**	**Одежда**	100 000
2000	Челябинск	**Косметика**	120 000
2000	**Челябинск**	[NULL]	**220 000**
2000	**Москва**	**Одежда**	250 000
2000	Москва	**Косметика**	75 000
2000	**Москва**	[NULL]	**325 000**
2000	[NULL]	[NULL]	**545 000**
2001	**Челябинск**	**Одежда**	230 000
2001	Челябинск	**Косметика**	310 000
2001	**Челябинск**	[NULL]	**540 000**
2001	**Москва**	**Одежда**	170 000
2001	Москва	**Косметика**	350 000
2001	**Москва**	[NULL]	**520 000**
2001	**[NULL]**	**[NULL]**	**1 060 000**
[NULL]	**[NULL]**	**[NULL]**	**1 605 000**

Корпоративные системы баз данных © М.Л. Цымблер 28

Анализ данных с помощью CUBE

```
select
   Время, Филиал,
   Товар, sum(Прибыль)
   as Прибыль
from Продажи
group by
    cube (Время,
          Филиал, Товар)
```

Время	Филиал	Товар	Прибыль
2000	**Челябинск**	**Одежда**	100 000
2000	Челябинск	**Косметика**	120 000
2000	**Челябинск**	**[NULL]**	**220 000**
2000	**Москва**	**Одежда**	250 000
2000	Москва	**Косметика**	75 000
2000	**Москва**	**[NULL]**	**325 000**
2000	[NULL]	**Одежда**	**350 000**
2000	[NULL]	**Косметика**	**195 000**
2000	**[NULL]**	**[NULL]**	**545 000**
2001	**Челябинск**	**Одежда**	230 000
2001	Челябинск	**Косметика**	310 000
2001	**Челябинск**	**[NULL]**	**540 000**
2001	**Москва**	**Одежда**	170 000
2001	Москва	**Косметика**	350 000
2001	**Москва**	**[NULL]**	**520 000**

Корпоративные системы баз данных © М.Л. Цымблер 29

Анализ данных с помощью CUBE

```
select
   Время, Филиал,
   Товар, sum(Прибыль)
   as Прибыль
from Продажи
group by
    cube (Время,
          Филиал, Товар)
```

Время	Филиал	Товар	Прибыль
[NULL]	**Челябинск**	**Одежда**	330 000
[NULL]	Челябинск	**Косметика**	430 000
[NULL]	**Челябинск**	**[NULL]**	**760 000**
[NULL]	**Москва**	**Одежда**	420 000
[NULL]	Москва	**Косметика**	425 000
[NULL]	**Москва**	**[NULL]**	**845 000**
[NULL]	[NULL]	**Одежда**	**750 000**
[NULL]	[NULL]	**Косметика**	**855 000**
[NULL]	**[NULL]**	**[NULL]**	**1 605 000**

Корпоративные системы баз данных © М.Л. Цымблер 30

Запросы Top-N

- Использование псевдо-столбца ROWNUM позволяет ограничить количество выводимых записей.

```
create view Лучшие_10_филиалов as
        select ROWNUM as Место, Филиал, Сумма
        from    (
                select Филиал, sum(Прибыль) as Сумма
                from Продажи
                group by Филиал
                order by sum(Прибыль) desc)
        where ROWNUM<=10;
```


Многомерная СУБД Oracle Express

- *Oracle Express Sever* – сервер многомерной СУБД на основе поликубической модели данных.
- *Oracle Express EDDiE* – утилита интерактивного администрирования многомерной базы данных (описание схемы, процедур загрузки и администрирования данных).
- *Oracle Express Analyzer* – система интерактивного построения OLAP-приложений, ориентированная на широкий круг пользователей (динамическое формирование запросов, отчетов, бизнес-диаграмм).
- *Oracle Express Objects* – система инструментальных средств, ориентированных на профессиональных разработчиков OLAP-приложений.

Определение данных в Oracle Express

```
-- Определение измерений
define Товар dimension text;
define Косметика dimension text;
define Обувь dimension text;
define Одежда dimension text;
define Год dimension year;
define Месяц dimension month;
define День dimension day;
define Регион dimension text;
define Город dimension text;
```

Определение данных в Oracle Express

```
-- Определение иерархии в измерениях
define Товар.Косметика
        relation Товар <Косметика>;
define Товар.Обувь
        relation Товар <Обувь>;
define Товар.Одежда
        relation Товар <Одежда>;
define Год.Месяц relation Год <Месяц>;
define Месяц.День relation Месяц <День>;
define Регион.Город relation Регион <Город>;
```


Определение данных в Oracle Express

```
-- Определение показателей
define Продажи
        variable decimal <День Товар Продавец>;
define Затраты
        variable decimal <День Товар Продавец>;
define Прибыль
        formula Продажи - Затраты decimal;
```


Манипулирование данными в Oracle Express

```
-- Вывод детализированного отчета
define Налог
        variable decimal temp <День Товар Продавец>;
define СтавкаНалога
        variable decimal temp;
set СтавкаНалога = 0.35;
limit День to Год 2003;
Налог = Прибыль * СтавкаНалога;
table Прибыль, Налог;
graph Прибыль, Налог;
```

Манипулирование данными в Oracle Express

```
-- Вывод агрегированного отчета
limit День to Месяц "Май";
limit День to Год 2003;
limit Город to "Челябинск";
limit Товар to "Косметика";
table total (Прибыль День);
```