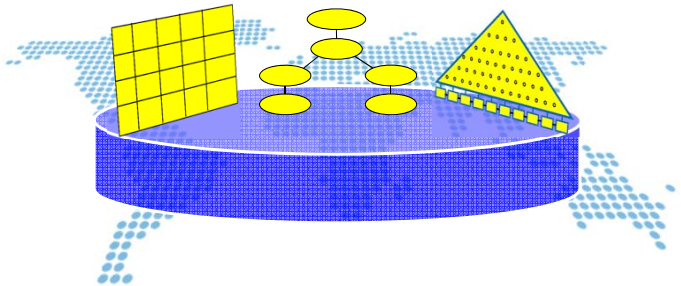
数据库系统

数据仓库

陈世敏

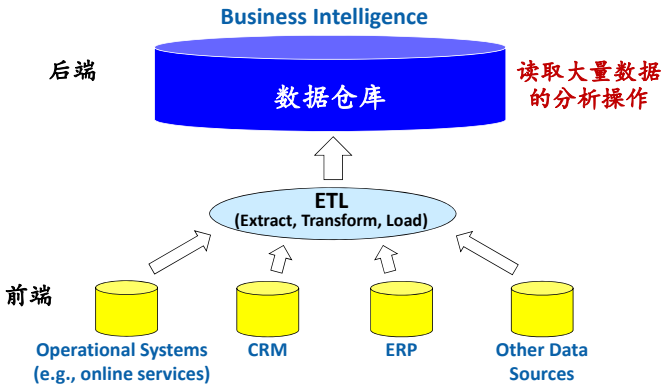
（中科院计算所）



Outline

- 数据仓库
- OLAP与Data Cube

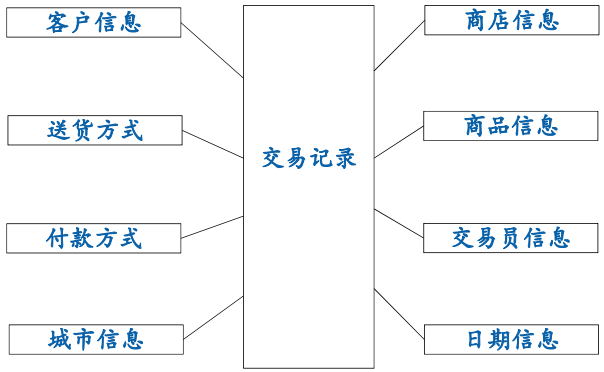
数据仓库 (Data Warehouse)



数据仓库 vs. 事务处理

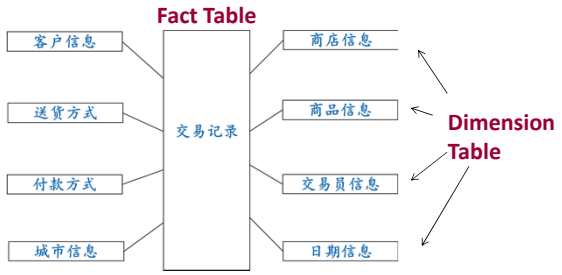
- 数据仓库
  - 少数数据分析操作
  - 每个操作访问大量的数据
  - 分析操作以读为主
- 事务处理
  - 大量的并发transactions
  - 每个transaction访问很少的数据
  - 读写

Star Schema：数据仓库中常见



Star Schema：数据仓库中常见

- 一个很大的fact table，多个dimension table
- Primary key – foreign key



常见的query形式

select …
from *fact*, *dim1*, *dim2*, …, *dimk*
where (*dim1.a op val1*) and
(*dim2.b op val2*) and
……
(*dimk.z op valk*) and
join key constraints
group by …
having …

在dimension表上施加约束条件，
对fact表进行统计分析

实现优化

- 列式存储（Column Store）
  - 如果只访问Fact表的少数几列
  - 可以提高效率
- 位图索引（Bitmap Index）
  - 可以有效地实现在多维上的过滤运算
- 固化视图（Materialized View）

列式数据存储

- 每个列产生一个文件，存储所有记录中该列的值

ID	Name	Birthday	Gender	Major	Year	GPA
131234	张飞	1995/1/1	男	计算机	2013	85
145678	貂蝉	1996/3/3	女	经管	2014	90
129012	孙权	1994/5/5	男	法律	2012	80
121101	关羽	1994/6/6	男	计算机	2012	90
142233	赵云	1996/7/7	男	计算机	2014	95

存储为7个列文件

ID
131234
145678
129012
121101
142233

Name
张飞
貂蝉
孙权
关羽
赵云

Birthday
1995/1/1
1996/3/3
1994/5/5
1994/6/6
1996/7/7

Gender
男
女
男
男
男

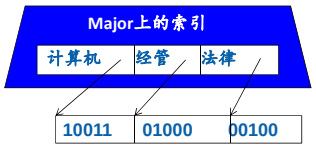
Major
计算机
经管
法律
计算机
计算机

Year
2013
2014
2012
2012
2014

GPA
85
90
80
90
95

Bitmap Index位图索引

ID	Name	Birthday	Gender	Major	Year	GPA
131234	张飞	1995/1/1	男	计算机	2013	85
145678	貂蝉	1996/3/3	女	经管	2014	90
129012	孙权	1994/5/5	男	法律	2012	80
121101	关羽	1994/6/6	男	计算机	2012	90
142233	赵云	1996/7/7	男	计算机	2014	95



- 每个不同的key对应一个bitmap
  - bitmap中的每位对应于一条记录
  - 1表示：这条记录在这列上取值==key
  - 0表示：这条记录在这列上取值!=key

利用位图索引计算数据仓库的查询

- select …
  from *fact*, *dim1*, *dim2*, …, *dimk*
  where (*dim1.a op val1*) and
  (*dim2.b op val2*) and
  ……
  (*dimk.z op valk*) and
  join key constraints
  group by …
  having …

- Fact表每个维度上面都建立位图索引

利用位图索引计算数据仓库的查询

- select …
  from *fact*, *dim1*, *dim2*, …, *dimk*
  where (*dim1.a op val1*) and
  (*dim2.b op val2*) and
  ……
  (*dimk.z op valk*) and
  join key constraints
  group by …
  having …

- 选择条件➔bitmap
- 用bitmap获取fact中相关记录

Materialized View（固化视图）

- 普通的View
- 固化视图
- 固化视图的更新

视图（View）

Student

ID	Name	Birthday	Gender	Major	Year	GPA
…	…	…	…	…	…	
…	…	…	…	…	…	

```
create view CS_Student as
      select *
       from   Student
       where Major='Computer Science';
```

- 实际上是一个被存下来的SQL语句
  - 记录create view信息，可能生成关系代数树
  - 但是，View的数据本身并没有提前计算
  - 在使用时，替换执行计划中相应部分

固化视图（Materialized View）

Student

ID	Name	Birthday	Gender	Major	Year	GPA
…	…	…	…	…	…	
…	…	…	…	…	…	

- 把视图的SQL语句执行了，结果放入了一个表
- 这样基于视图的运算可以无需重复计算

```
create materialized view CS_Student as
      select *
       from   Student
       where Major='Computer Science';
```


固化视图的更新

- 随着时间推移，基础表中数据不断增加和变化
- 需要更新固化视图来反映基础表的变化
- 方式1：重新计算
  - 例如：PostgreSQL的refresh操作
  - 通常手工进行
  - 代价高，用时长

固化视图的更新

- 方式2：增量计算
  - 目标是避免重复计算
    - 计算量与delta数据成正比，而不是与全部数据量成正比
  - 1990~2000年的很多研究工作，有比较成熟的理论
    - 对典型的SQL关系运算都进行了研究
    - T=R⋈S，那么：ΔT= (ΔR ⋈S)∪(R ⋈ΔS))∪(ΔR ⋈ ΔS)
    - 哪些运算可以高效支持？哪些操作无法增量计算？
      - 例如，选择，投影，连接，sum, avg, count都可以支持
      - 但是如果有delete/update，max, min就很难支持
  - 在开源系统中支持较差
  - 商业数据库中，例如Oracle，支持固化视图的增量计算

Outline

- 数据仓库简介
- OLAP与Data Cube

OLAP

- Online Analytical Processing（联机分析处理）
- 数据仓库通常是OLAP的基础
  - OLAP是在数据仓库的基础上实现的
- OLAP的基本数据模型是多维矩阵
  - 例如，在多个dimension上进行group by操作
  - 得到的多维矩阵的每项代表一个分组，
    每项的值是Fact表上对于这个分组的聚集统计值
- 称作：Data Cube（数据立方）

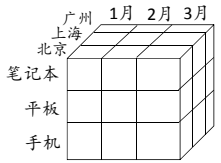
Data Cube(数据立方)

- 例如，二维的数据立方，记录分组的统计数据

时间

商品	1月	2月	3月
笔记本	1000	1500	1600
平板	2000	2500	3000
手机	3000	3100	3200

- 例如，三维的数据立方



Data Cube(数据立方)

- 多维的数据表示
  - 适合对趋势的分析
  - 可以从宏观到微观，从微观到宏观
- 常用操作
  - Roll up / drill down
  - Slice, dice

Data Cube(数据立方)：rollup（上卷）

- 例如，二维的数据立方

Rollup 时间维度

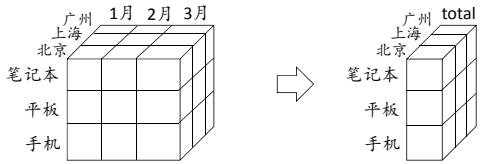
时间

商品	1月	2月	3月
笔记本	1000	1500	1600
平板	2000	2500	3000
手机	3000	3100	3200

⇨

商品	total
笔记本	4100
平板	7500
手机	9300

- 例如，三维的数据立方



Data Cube(数据立方)：drill down(下钻)

- 例如，二维的数据立方

时间维度

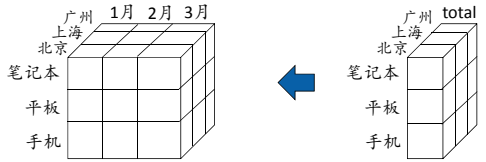
时间

商品	1月	2月	3月
笔记本	1000	1500	1600
平板	2000	2500	3000
手机	3000	3100	3200

⇦

商品	total
笔记本	4100
平板	7500
手机	9300

- 例如，三维的数据立方



概念层级

- 在一个维度上有可能可以定义层级
  - 时间：年-月-日
  - 地点：国家-省-市
  - 商品：品种-具体型号-不同厂家的同类产品
  - 等等
- 前面的例子
  - Roll up：在某维上求和，降维
  - Drill down：把某维的和分解，增维
- 还可以对概念层级操作
  - Roll up：在某维上，从细粒度到粗粒度
  - Drill down：在某维上，从粗粒度到细粒度

Data Cube(数据立方)：slice(切片)

在某维上选一个值

- 例如，二维的数据立方

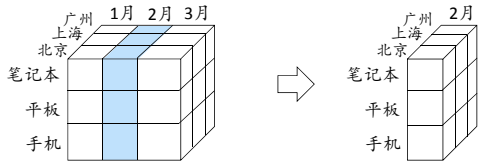
时间

商品	1月	2月	3月
笔记本	1000	1500	1600
平板	2000	2500	3000
手机	3000	3100	3200

⇨

商品	2月
笔记本	1500
平板	2500
手机	3100

- 例如，三维的数据立方



Data Cube(数据立方)：dice（切块）

在多维上选多个值

- 例如，二维的数据立方

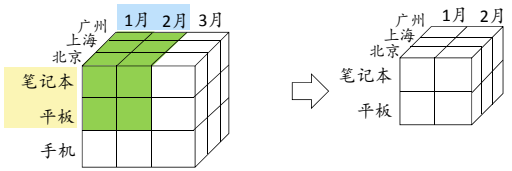
时间

商品	1月	2月	3月
笔记本	1000	1500	1600
平板	2000	2500	3000
手机	3000	3100	3200

⇨

商品	1月	2月
笔记本	1000	1500
平板	2000	2500

- 例如，三维的数据立方



ROLAP和MOLAP

- ROLAP（Relational OLAP）
  - 数据存储在关系表中，例如star schema
  - 采用查询语言、索引结构等来支持OLAP的运算
- MOLAP（Multidimensional OLAP）
  - 计算得到最基础的信息后，从数据库获取数据
  - 采用专门的多维数据结构来表达Data Cube
  - 在Data Cube上进行操作

Group by … with cube

```
create materialized view StudentCube as
  select major, year, COUNT(*)
  from Student
  group by major, year with cube;
```

Student

ID	Name	Birthday	Gender	Major	Year	GPA
131234	张飞	1995/1/1	男	计算机	2013	85
145678	貂蝉	1996/3/3	女	经管	2014	90
129012	孙权	1994/5/5	男	法律	2012	80
121101	关羽	1994/6/6	男	计算机	2012	90
142233	赵云	1996/7/7	男	计算机	2014	95

让我们看一下普通的group by的结果

```
select major, year, COUNT(*)
  from Student
  group by  major, year;
```

Major	Year	count(*)
计算机	2012	1
计算机	2013	1
计算机	2014	1
经管	2014	1
法律	2012	1

Student

ID	Name	Birthday	Gender	Major	Year	GPA
131234	张飞	1995/1/1	男	计算机	2013	85
145678	貂蝉	1996/3/3	女	经管	2014	90
129012	孙权	1994/5/5	男	法律	2012	80
121101	关羽	1994/6/6	男	计算机	2012	90
142233	赵云	1996/7/7	男	计算机	2014	95

with cube的结果

```
select major, year, COUNT(*)
  from Student
  group by  major, year
     with cube;
```

逻辑上cube是这样的

Major	2012	2013	2014	total
计算机	1	1	1	3
经管			1	1
法律	1			1
total	2	1	2	5

输出结果具体表达如下

Major	Year	count(*)
计算机	2012	1
计算机	2013	1
计算机	2014	1
经管	2014	1
法律	2012	1
计算机	NULL	3
经管	NULL	1
法律	NULL	1
NULL	2012	2
NULL	2013	1
NULL	2014	2
NULL	NULL	5

小结

- 数据仓库简介
- OLAP与Data Cube