Data Warehousing
Introduction

Toon Calders

toon.calders@ulb.ac.be
Course Organization

• Lectures on Tuesday 14:00 and Friday 16:00
  – Check [http://gehol.ulb.ac.be/](http://gehol.ulb.ac.be/) for room

• Most exercises in computer class
  – Tutorial MS SQL Server tools
    • MS Sequel Server, SSIS, SSAS, SSRS

• Contributions from associated partners
  – IBM (TBC)
  – Teradata (TBC)
Course Organization

• Grading:
  – Written exam (14/20)
  – Project (6/20)

• 4 practical assignments in groups of 3-4
  – Dimensional modeling, logical model,
  – Initial load
  – ETL for Data Warehouse Update; changing dimensions
  – Data cube construction & reports
Motivation for the Course

• Database = a piece of software to handle data:
  – Store, maintain, and query

• Most ideal system situation-dependent
  – data type: simple / semi-structured / complex / ... 
  – types of queries: simple lookup / analytical / ... 
  – type of usage: multi-user / single-user / distributed / ...
Online Transaction Processing (OLTP)

• Relational database management systems are mainly to support transaction processing
  – Concurrent access
  – Data consistency, non-redundancy
  – Ad-hoc Querying
  – Efficiency
Atomicity

• Consider a Bank transaction; John transfers 100 euro to Mary
  1. Check if Balance John > 100 euro?
  2. Balance John -100 euro
  3. Balance Mary +100 euro

• What can go wrong when the banking system crashes?
Atomicity

• Consider a Bank transaction; John transfers 100 euro to Mary
  1. Check if Balance John > 100 euro?
  2. Balance John -100 euro
  3. Balance Mary +100 euro

• What can go wrong when the banking system crashes?
  – When the system is restarted, John has 100 euro less, but Mary did not receive it!
Consistency

• Consider a Bank transaction; John transfers 100 euro to Mary
  1. Balance John -100 euro
  2. Balance Mary +100 euro

• Suppose consistency rule:
  Balance should always $\geq 0$
  – After the transaction, the database should still be consistent
  – Otherwise: roll-back
Durability

• Consider a Bank transaction; John transfers 100 euro to Mary
  1. Check if Balance John > 100 euro?
  2. Balance John -100 euro
  3. Balance Mary +100 euro

COMMIT

CRASH

• After commit, transaction result should persist
Isolation

- Consider a Bank transaction; John withdraws 100 euro from an ATM; his wife Mary pays 50 Euro in a shop at the same time, from the same account.

  John
  Get balance
  Subtract 100 euro
  Store new balance

  Mary
  Get balance
  Subtract 50 euro
  Store new balance

- Possible problems?
Isolation

- Consider a Bank transaction; John withdraws 100 euro from an ATM; his wife Mary pays 50 Euro in a shop at the same time, from the same account.

<table>
<thead>
<tr>
<th>John</th>
<th>Mary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Get balance</td>
<td>1b. Get balance</td>
</tr>
<tr>
<td>2a. Subtract 100 euro</td>
<td>2b. Subtract 50 euro</td>
</tr>
<tr>
<td>3a. Store new balance</td>
<td>3b. Store new balance</td>
</tr>
</tbody>
</table>
Isolation

• Consider a Bank transaction; John withdraws 100 euro from an ATM; his wife Mary pays 50 Euro in a shop at the same time, from the same account.

  John
  1a. Get balance
  2a. Subtract 100 euro
  3a. Store new balance

  Mary
  1b. Get balance
  2b. Subtract 50 euro
  3b. Store new balance

-150
-100
-50
Concurrent Access

• Multiple users
  – Concurrent access
  – Frequent inserts, deletes, updates
  → need for ACID

• Extremely important to have most recent information

• Enforced by “protocols” based on *locking*
Online Transaction Processing (OLTP)

- Relational database management systems are mainly to support transaction processing
  - Concurrent access
  - Data consistency, non-redundancy
  - Ad-hoc Querying
  - Efficiency
Design Theory

• Which instance do you prefer? Why?

<table>
<thead>
<tr>
<th>Student</th>
<th>Code</th>
<th>Name</th>
<th>Semester</th>
<th>Lecturer</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phil</td>
<td>2ID45</td>
<td>Advanced Databases</td>
<td>Spring 2011</td>
<td>Calders</td>
<td>A+</td>
</tr>
<tr>
<td>Mary</td>
<td>2ID45</td>
<td>Advanced Databases</td>
<td>Spring 2011</td>
<td>Calders</td>
<td>C</td>
</tr>
<tr>
<td>John</td>
<td>2ID45</td>
<td>Advanced Databases</td>
<td>Spring 2011</td>
<td>Calders</td>
<td>B-</td>
</tr>
<tr>
<td>Paul</td>
<td>2ID05</td>
<td>Databases I</td>
<td>Spring 2011</td>
<td>Fletcher</td>
<td>C</td>
</tr>
</tbody>
</table>

Courses

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2ID45</td>
<td>Advanced Databases</td>
</tr>
<tr>
<td>2ID05</td>
<td>Databases I</td>
</tr>
</tbody>
</table>

Offerings

<table>
<thead>
<tr>
<th>Code</th>
<th>Semester</th>
<th>Lecturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2ID45</td>
<td>Spring 2011</td>
<td>Calders</td>
</tr>
<tr>
<td>2ID05</td>
<td>Spring 2011</td>
<td>Fletcher</td>
</tr>
</tbody>
</table>
Revisiting Relational Database

Conceptual Model → Relational Model → Database

SQL querying
• Models entities and relations between them
  – “language” to write down constraints
  – documentation of the database design
Relational Model

• Relational Databases store the data in tables
  patient(SSN,fname,sname,city,street,age)
  doctor(ID,name,speciality,experience)
  treatment(SSN,Nr,ID,cost,start,end)
  drug(code,name,type,price)
  includes(SSN,Nr,code,quantity)
  room(rnr,capacity,price)
  stay(SSN,rnr,from,to)

• Good design =
  – No redundancy ➔ limit danger of inconsistencies
  – Constraints as much as possible covered by the design of the tables
Online Transaction Processing (OLTP)

- Relational database management systems are mainly to support transaction processing
  - Concurrent access
  - Data consistency, non-redundancy
  - Ad-hoc Querying
  - Efficiency
Powerful Language SQL

• Ad-hoc querying

```sql
SELECT fname, sname
FROM Customer
WHERE SSN="778944";
```

```sql
SELECT distinct S.name
FROM supplier S, transaction T, customer C
WHERE C.city="Brussels"
and S.SID=T.SID and C.SSN=T.SSN;
```

```sql
SELECT S.City, sum(T.price), avg(T.price)
FROM supplier S, transaction T, customer C
WHERE C.city="Brussels"
and S.SID=T.SID and C.SSN=T.SSN
GROUP BY S.City;
```
General-Purpose Language SQL

- Database engine optimizes queries
  - Makes a *query plan*
  - Using database statistics

- General rule of thumb:

  *The more powerful the query language, the more difficult it is to automatically optimize it*
Online Transaction Processing (OLTP)

• Relational database management systems are mainly to support transaction processing
  – Concurrent access
  – Data consistency, non-redundancy
  – Ad-hoc Querying
  – Efficiency
Indexing Principle

No index
Indexing Principle

• Database Equivalent

No index ➔ Expensive

Full table scan

Inexpensive

index lookup

+ Retrieve data page
Summary: Relational DBMS

- Strong in supporting OLTP
- Mainly aimed towards many, frequent, concurrent, small, ad-hoc queries
What About Decision Support?

Decision support
- Off-line setting
- « Historical » data
- Summarized data
- Integrate different databases
- Statistical queries

Flight company
- Evaluate ROI flights
- Flights of last year
- # passengers per carrier for destination X
- Passengers, fuel costs, maintenance info
- Average % of seats sold/month/destination
Create Reports

### Sales Summary

<table>
<thead>
<tr>
<th>Sales Manager</th>
<th>Details</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2004 Sales Over Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jien, Stephen</td>
<td>Pak, Jae</td>
<td>20,330,657</td>
<td>22,930,001</td>
<td>10,931,024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Varkey, Ranjit</td>
<td>2,522,836</td>
<td>4,172,459</td>
<td>1,808,043</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valdez, Rachel</td>
<td>848,333</td>
<td>2,286,700</td>
<td>1,374,856</td>
<td></td>
</tr>
<tr>
<td>Abbas, Syed</td>
<td></td>
<td>978,435</td>
<td>848,632</td>
<td></td>
<td>On Track</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3,371,169</td>
<td>7,437,594</td>
<td>4,031,531</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>701,487</td>
<td>720,324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>23,701,827</td>
<td>31,069,082</td>
<td>15,682,879</td>
<td></td>
</tr>
</tbody>
</table>

### Percent returns by state

- 0% - 11%
- 11% - 22%
- 22% - 33%
- 33% - 44%

### Total sales by state (in thousands)

- 130
- 380
- 850
- 1,200
- 1,900

- 20
- 270
- 490
- 1,100
- 1,700
- 2,300
### Browse Data

<table>
<thead>
<tr>
<th>Region</th>
<th>Unit Sales</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Stores</td>
<td>509,987</td>
<td>137,078</td>
<td>135,745</td>
<td>139,412</td>
<td>97,752</td>
</tr>
<tr>
<td>Store Sales</td>
<td>1,079,147.47</td>
<td>290,873.18</td>
<td>287,009.99</td>
<td>295,040.55</td>
<td>206,223.75</td>
</tr>
<tr>
<td>Canada</td>
<td>46,157</td>
<td>11,160</td>
<td>12,885</td>
<td>12,966</td>
<td>9,146</td>
</tr>
<tr>
<td>Store Sales</td>
<td>98,045.46</td>
<td>23,881.13</td>
<td>27,685.00</td>
<td>27,176.30</td>
<td>19,303.03</td>
</tr>
<tr>
<td>BC</td>
<td>46,157</td>
<td>11,160</td>
<td>12,885</td>
<td>12,966</td>
<td>9,146</td>
</tr>
<tr>
<td>Store Sales</td>
<td>98,045.46</td>
<td>23,881.13</td>
<td>27,685.00</td>
<td>27,176.30</td>
<td>19,303.03</td>
</tr>
<tr>
<td>Mexico</td>
<td>203,914</td>
<td>56,133</td>
<td>54,005</td>
<td>57,872</td>
<td>35,904</td>
</tr>
<tr>
<td>Store Sales</td>
<td>430,293.59</td>
<td>118,589.41</td>
<td>113,830.59</td>
<td>122,706.05</td>
<td>75,167.54</td>
</tr>
<tr>
<td>DF</td>
<td>45,223</td>
<td>12,058</td>
<td>12,818</td>
<td>12,962</td>
<td>7,385</td>
</tr>
<tr>
<td>Store Sales</td>
<td>95,526.40</td>
<td>25,590.39</td>
<td>27,096.37</td>
<td>27,350.86</td>
<td>15,488.78</td>
</tr>
<tr>
<td>Guerrero</td>
<td>23,226</td>
<td>7,042</td>
<td>5,885</td>
<td>6,008</td>
<td>4,291</td>
</tr>
<tr>
<td>Store Sales</td>
<td>49,090.03</td>
<td>15,063.14</td>
<td>12,301.53</td>
<td>12,755.76</td>
<td>8,969.60</td>
</tr>
<tr>
<td>Jalisco</td>
<td>2,124</td>
<td>666</td>
<td>637</td>
<td>492</td>
<td>329</td>
</tr>
<tr>
<td>Store Sales</td>
<td>4,328.87</td>
<td>1,356.81</td>
<td>1,246.77</td>
<td>1,035.42</td>
<td>689.87</td>
</tr>
<tr>
<td>Veracruz</td>
<td>24,696</td>
<td>6,711</td>
<td>6,119</td>
<td>6,947</td>
<td>4,919</td>
</tr>
<tr>
<td>Store Sales</td>
<td>52,142.07</td>
<td>13,970.82</td>
<td>13,114.47</td>
<td>14,727.55</td>
<td>10,329.23</td>
</tr>
<tr>
<td>Yucatan</td>
<td>37,143</td>
<td>9,766</td>
<td>9,372</td>
<td>11,205</td>
<td>6,800</td>
</tr>
<tr>
<td>Store Sales</td>
<td>79,063.13</td>
<td>20,592.65</td>
<td>19,909.69</td>
<td>24,247.97</td>
<td>14,312.82</td>
</tr>
<tr>
<td>Zacatecas</td>
<td>71,502</td>
<td>19,890</td>
<td>19,174</td>
<td>20,258</td>
<td>12,180</td>
</tr>
<tr>
<td>Store Sales</td>
<td>150,143.09</td>
<td>42,015.60</td>
<td>40,161.76</td>
<td>42,588.49</td>
<td>25,377.24</td>
</tr>
<tr>
<td>USA</td>
<td>259,916</td>
<td>69,785</td>
<td>68,855</td>
<td>68,574</td>
<td>52,702</td>
</tr>
<tr>
<td>Store Sales</td>
<td>550,808.42</td>
<td>148,402.64</td>
<td>145,494.40</td>
<td>145,158.20</td>
<td>111,753.18</td>
</tr>
</tbody>
</table>
Example: Business Case

• Company selling different products
  – “units” of a high-tech material
  – different parameters
  – base product for other (high-tech) products
  – B2B scenario

• Company sees profit is dropping
  – Why?
Example: Business Case

• Different salesmen sell the products to their customers
  – Different price; result of negotiation
  – Transaction stored in sales database
    • Some transactions are to “compensate” incorrect transactions
  – There are seasonal effects (less sales in winter)
  – Data spread over different branches; formats are slightly different
### Example: Business Case

#### Example Inc., August 2012

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>4,237</td>
<td>32,916</td>
<td>3,987</td>
<td>53,000</td>
<td>49,374</td>
<td>52,000</td>
<td>1,000</td>
<td>3,626</td>
<td></td>
</tr>
<tr>
<td>Total sales</td>
<td>4,237</td>
<td>32,916</td>
<td>3,987</td>
<td>53,000</td>
<td>49,374</td>
<td>52,000</td>
<td>1,000</td>
<td>3,626</td>
<td></td>
</tr>
<tr>
<td>Costs of goods sold</td>
<td>1,983</td>
<td>15,405</td>
<td>1,866</td>
<td>24,804</td>
<td>23,107</td>
<td>24,336</td>
<td>468</td>
<td>1,697</td>
<td>Standard %</td>
</tr>
<tr>
<td>% of total sales</td>
<td>46.8%</td>
<td>46.8%</td>
<td>46.8%</td>
<td>46.8%</td>
<td>46.8%</td>
<td>46.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribution cost</td>
<td>1,215</td>
<td>9,612</td>
<td>998</td>
<td>13,875</td>
<td>14,418</td>
<td>15,000</td>
<td>-1,125</td>
<td>-543</td>
<td></td>
</tr>
<tr>
<td>% of total sales</td>
<td>28.7%</td>
<td>29.2%</td>
<td>25.0%</td>
<td>26.2%</td>
<td>29.2%</td>
<td>28.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gross margin</td>
<td>1,039</td>
<td>7,899</td>
<td>1,123</td>
<td>14,321</td>
<td>11,849</td>
<td>12,664</td>
<td>1,657</td>
<td>2,472</td>
<td></td>
</tr>
<tr>
<td>% of total sales</td>
<td>24.5%</td>
<td>24.0%</td>
<td>28.2%</td>
<td>27.0%</td>
<td>24.0%</td>
<td>24.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expenses</td>
<td>214</td>
<td>1,712</td>
<td>211</td>
<td>2,568</td>
<td>2,568</td>
<td>2,568</td>
<td>0</td>
<td>0</td>
<td>Fixed</td>
</tr>
<tr>
<td>% of total sales</td>
<td>5.1%</td>
<td>5.2%</td>
<td>5.3%</td>
<td>4.8%</td>
<td>5.2%</td>
<td>4.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admin</td>
<td>115</td>
<td>920</td>
<td>112</td>
<td>1,380</td>
<td>1,380</td>
<td>1,380</td>
<td>0</td>
<td>0</td>
<td>Fixed</td>
</tr>
<tr>
<td>% of total sales</td>
<td>2.7%</td>
<td>2.8%</td>
<td>2.8%</td>
<td>2.6%</td>
<td>2.8%</td>
<td>2.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R&amp;D</td>
<td>36</td>
<td>312</td>
<td>42</td>
<td>465</td>
<td>468</td>
<td>465</td>
<td>0</td>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>% of total sales</td>
<td>0.8%</td>
<td>0.9%</td>
<td>1.1%</td>
<td>0.9%</td>
<td>0.9%</td>
<td>0.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>EBITA</td>
<td>674</td>
<td>4,955</td>
<td>758</td>
<td>9,908</td>
<td>7,433</td>
<td>8,251</td>
<td>1,657</td>
<td>2,475</td>
<td></td>
</tr>
<tr>
<td>% of total sales</td>
<td>15.9%</td>
<td>15.1%</td>
<td>19.0%</td>
<td>18.7%</td>
<td>15.1%</td>
<td>15.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depreciation</td>
<td>410</td>
<td>3,280</td>
<td>410</td>
<td>4,920</td>
<td>4,920</td>
<td>4,920</td>
<td>0</td>
<td>0</td>
<td>Fixed</td>
</tr>
<tr>
<td>% of total sales</td>
<td>9.7%</td>
<td>10.0%</td>
<td>10.3%</td>
<td>9.3%</td>
<td>10.0%</td>
<td>9.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBITDA</td>
<td>264</td>
<td>1,675</td>
<td>348</td>
<td>4,988</td>
<td>2,513</td>
<td>3,331</td>
<td>1,657</td>
<td>2,475</td>
<td></td>
</tr>
<tr>
<td>% of total sales</td>
<td>6.2%</td>
<td>5.1%</td>
<td>8.7%</td>
<td>9.4%</td>
<td>5.1%</td>
<td>6.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: Business Case

• Gathering the sales data took considerable time
• Data needed to be cleaned
• Analysis questions
  – Average, minimal, maximal price per region/salesman for comparable transactions
  – Average sales per product type and region
  – Evolution of sales this year over time, compared to last year’s sales
Example: Business Case

• Typically: want to browse the data
  – Explore
  – Concentrate on certain slices of the data
  – Refine analysis in a suspicious region
  – ...

• Almost impossible using original data sources and OLTP-geared systems
Requirements for Decision Support?

• Concurrent access
  → not really
  → read-only

• Data consistency, non-redundancy
  → data comes from consistent sources (sort of)
  → data does not change during analysis; once clean, always clean
Requirements for Decision Support?

- Ad-hoc Querying
  - No longer true;
  - Spread-sheet like queries
  - Long-running queries, touching large parts of the database
    - In combination with transactions, kills the database

- Efficiency
  - Relational DBMS optimized for other types of queries
Requirements for Decision Support?

• OLTP systems not very efficient for data analysis tasks
  – analysis queries might stall operational systems
  – architecture suboptimal
    • different indexing structures
    • denormalization
  – need of historical data versus only current data
Outline

Online Analytical Processing

• **Data Warehouses**
• Conceptual model: Data Cubes
• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX
• Database Explosion Problem
Data Warehouse

• A decision support DB maintained separately from the operational databases.

• Why Separate Data Warehouse?
  – Different functions
    • DBMS—tuned for OLTP
    • Warehouse—tuned for OLAP
  – Different data
    • Decision support requires historical data
  – Integration of data from heterogeneous sources
Data Warehouse

• Data Warehouse is
  – Subject-oriented (vs function-oriented)
  – Non-volatile (vs only holding most recent version)
  – Integrated (different data sources)
  – Time-variant (can be related to time)
  – Supporting decision support
Three-Tier Architecture

- **Data Sources**
  - OLTP systems (many users)
  - other sources
  - Operational DBs
  - Oracle, IBM DBII, MS SQL Server, MySQL, PostgreSQL

- **Data Storage**
  - Data Warehouse: SSIS, IBM DataStage, Informatica, TALEND, ...
  - ROLAP Server: Analysis
  - OLAP Server: Query/Reporting
  - Mondrian: Data Mining

- **OLAP Engine**
  - SSAS
  - OLAP

- **Front-End Tools**
  - Microstrategy, tableau, SSRS TARGIT, ...
  - Analysis
  - Query/Reporting

- **Extract, Transform, Load, Refresh**
  - Extract
  - Transform
  - Load
  - Refresh

- **Data Marts**
  - Operational DBs
  - other sources
  - Oracle, IBM DBII, MS SQL Server, MySQL, PostgreSQL

- **OLTP systems**
Example: MS SQLSERVER

SSIS: SQL Server Integration Services
SSAS: SQL Server Analysis Services
SSRS: SQL Server Reporting Services

Operational Setting

Client

OLTP

SQL Server

SSIS

ETL

SQL Server

ROLAP

SSAS

MOLAP

SSRS

report

Browse cube
Example: Top-Down

Operational DBs

SQL Server
Oracle

Staging area

ETL

DBMS

Data Warehouse

ETL

Data mart
Data mart
Data mart

Log file
XML doc.
Example: Bottom-Up

Staging area

Operational DBs
- SQL Server
- Oracle

ETL

Data mart

MDDB

Data Warehouse

Ralph Kimball
OLAP

• OLAP = OnLine Analytical Processing
  – Online = no waiting for answers

• OLAP system = system that supports *analytical queries* that are *dimensional* in nature.

• Most data warehousing systems support OLAP functionalities
Outline

Online Analytical Processing

• Conceptual model: Data cubes

• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX

• Database Explosion Problem
Supermarket Example

• Evaluate the sales of products
  - **Product cost in $**
  - **Customer:** ID, city, state, country,
    - **Store:** chain, size, location,
    - **Product:** brand, type, ...
  - ...

![Diagram showing dimensions and hierarchies in supermarket example](image-url)
Supermarket Example

• Multi-dimensional view on data
Cross Tabulation

- Cross-tabulations are highly useful
  - Sales of clothes June → August ‘06

<table>
<thead>
<tr>
<th></th>
<th>Blue</th>
<th>Red</th>
<th>Orange</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>June</td>
<td>51</td>
<td>25</td>
<td>158</td>
<td>234</td>
</tr>
<tr>
<td>July</td>
<td>58</td>
<td>20</td>
<td>120</td>
<td>198</td>
</tr>
<tr>
<td>August</td>
<td>65</td>
<td>22</td>
<td>51</td>
<td>138</td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
<td>67</td>
<td>329</td>
<td>570</td>
</tr>
</tbody>
</table>

Date: month, June → August 2006
Data Cubes

- Extension of Cross-Tables to multiple dimensions
  - *Conceptual* notion

![Data Cube Diagram]

<table>
<thead>
<tr>
<th></th>
<th>Blue</th>
<th>Red</th>
<th>Orange</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>June</td>
<td>51</td>
<td>25</td>
<td>158</td>
<td>234</td>
</tr>
<tr>
<td>July</td>
<td>58</td>
<td>20</td>
<td>120</td>
<td>198</td>
</tr>
<tr>
<td>August</td>
<td>65</td>
<td>22</td>
<td>51</td>
<td>138</td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
<td>67</td>
<td>329</td>
<td>570</td>
</tr>
</tbody>
</table>

Dimensions

- Aggregated w.r.t. X-dim
- Aggregated w.r.t. Y-dim
- Aggregated w.r.t. X and Y
### Data Cubes

The diagram illustrates a data cube structure with dimensions for **Product**, **Date**, and **Country**. The cube is segmented by **TV**, **PC**, and **VCR** products, with **1Qtr**, **2Qtr**, **3Qtr**, and **4Qtr** dates, and **Ireland**, **France**, and **Germany** countries. The **sum** values are aggregated over **country**, **date**, and **product**.
Data Cubes

Aggregated over
- product & date
- date & country
- product & country
Data Cubes

- **Product**: TV, PC, VCR
- **Country**: Ireland, France, Germany
- **Date**: 1Qtr, 2Qtr, 3Qtr, 4Qtr, sum

Summarized over all regions.
SSAS – Data Cubes

Cube name

Measures

“key performance indicators”

Dimensions
SSAS - Dimension

- Customer
  - Demographic
    - Commute Distance
    - Education
    - Gender
    - Home Owner
    - Marital Status
    - Number of Cars Owned
    - Number of Children At Home
    - Occupation
    - Total Children
    - Yearly Income
  - Location
    - City
    - Country
    - Postal Code
    - State-Province
  - Sets
  - Customer
  - Customer Geography
    - Members
    - Country
    - State-Province
    - City
    - Postal Code
    - Customer

attributes

Hierarchy

Levels
Hierarchy, Level

- One dimension can have multiple hierarchies
- Hierarchies consist of \textit{levels}
- Levels are in a linear order
# Member

## Hierarchy Geography on Dimension Customer

<table>
<thead>
<tr>
<th>ALL</th>
<th>Country</th>
<th>State-Province</th>
<th>City</th>
<th>Postal code</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Australia</td>
<td>New South Wales</td>
<td>Alexandria</td>
<td>2015</td>
<td>Adriana Smith</td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>Queensland</td>
<td>Coffs Harbour</td>
<td>2450</td>
<td>Aimee Guo</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>South Australia</td>
<td>Darlinghurst</td>
<td>2010</td>
<td>Allison R. Young</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>Tasmania</td>
<td>Goulburn</td>
<td>2580</td>
<td>Ann A. Sara</td>
</tr>
<tr>
<td></td>
<td>United Kingdom</td>
<td>Victoria</td>
<td>Lane Cove</td>
<td>1597</td>
<td>Antonio G. Patterson</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>Alberta</td>
<td>Lavender Bay</td>
<td>2060</td>
<td>Ariana Stewart</td>
</tr>
<tr>
<td></td>
<td></td>
<td>British Columbia</td>
<td>Malabar</td>
<td>2036</td>
<td>Arthur Kapoor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brunswick</td>
<td>Matraville</td>
<td>2036</td>
<td>Barbara W. Lal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manitoba</td>
<td>Milsons Point</td>
<td>2036</td>
<td>Bobby D. Saunders</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ontario</td>
<td>Newcastle</td>
<td>2061</td>
<td>Brianna J. Johnson</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quebec</td>
<td>North Ryde</td>
<td>2300</td>
<td>Bruce G. Madan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Charente-Maritime</td>
<td>North Sydney</td>
<td>2113</td>
<td>Bryant L. Perez</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Essonne</td>
<td>Port Macquarie</td>
<td>2055</td>
<td>Carla D. Madan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Garonne (Haute)</td>
<td>Rhodes</td>
<td>2444</td>
<td>Carlos Edwards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gers</td>
<td>Silverwater</td>
<td>2138</td>
<td>Carly Anand</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hauts de Seine</td>
<td>Springwood</td>
<td>2264</td>
<td>Cedric Liu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loir et Cher</td>
<td>St. Leonards</td>
<td>2777</td>
<td>Clarence Xu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sydney</td>
<td>2065</td>
<td>Colin Chavez</td>
</tr>
</tbody>
</table>
Children, Parent

Hierarchy Geography on Dimension Customer

<table>
<thead>
<tr>
<th>ALL</th>
<th>Country</th>
<th>State-Province</th>
<th>City</th>
<th>Postal code</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Members

- Australia
- Canada
- France
- Germany
- United Kingdom
- United States

Children

- New South Wales
- Queensland
- South Australia
- Tasmania
- Victoria
- Alberta
- British Columbia
- Nova Scotia

Parents

- Alexandria
- Coffs Harbour
- Darlinghurst
- Goulburn
- Lane Cove
- Lavender Bay
- Malabar
- Matraville
- Milsons Point
- Newcastle
- North Ryde
- North Sydney
- Port Macquarie
- Rhodes
- Silverwater
- Springwood
- St. Leonards
- Sydney

Postal code

- 2015
- 2450
- 2010
- 2580
- 1597
- 2060
- 2036
- 2036
- 2061
- 2300
- 2113
- 2055
- 2444
- 2138
- 2264
- 2777
- 2065
- 1002

Members

- Adriana Smith
- Aimee Guo
- Allison R. Young
- Ann A. Sara
- Antonio G. Patterson
- Ariana Stewart
- Arthur Kapoor
- Barbara W. Lal
- Bobby D. Saunders
- Brianna J. Johnson
- Bruce G. Madan
- Bryant L. Perez
- Carla D. Madan
- Carlos Edwards
- Carly Anand
- Cedric Liu
- Clarence Xu
- Colin Chavez

...