Notice

- Author
  - João Moura Pires (jmp@di.fct.unl.pt)

- This material can be freely used for personal or academic purposes without any previous authorization from the author, only if this notice is maintained with.

- For commercial purposes the use of any part of this material requires the previous authorization from the author.
Bibliography

Many examples are extracted and adapted from


Table of Contents

- Corporate Information Factory

- Quick overview of OLAP cube concepts

- Basics of Multidimensional Modeling
Corporate Information Factory

Strategic and tactical portions of a BI environment.

[Imhoff, 2003]
Data acquisition is a set of processes and programs that extracts data for the data warehouse and operational data store from the operational systems. The data acquisition programs perform the cleansing as well as the integration of the data and transformation into an enterprise format.

This enterprise format reflects an integrated set of enterprise business rules that usually causes the data acquisition layer to be the most complex component in the CIF. In addition to programs that transform and clean up data, the data acquisition layer also includes audit and control processes and programs to ensure the integrity of the data as it enters the data warehouse or operational data store.
Data delivery is the process that moves data from the data warehouse into data and operation marts. Like the data acquisition layer, it manipulates the data as it moves it. In the case of data delivery, however, the origin is the data warehouse or ODS, which already contains high-quality, integrated data that conforms to the enterprise business rules.

“a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making” [Immon, 1980]
CIF: Operational Data Store

- It is subject oriented like a data warehouse.
- Its data is fully integrated like a data warehouse.
- Its data is current.
  The ODS has minimal history and shows the state of the entity as close to real time as feasible.
- Its data is volatile or updatable.
- Its data is almost entirely detailed with a small amount of dynamic aggregation.

CIF: Data Mart

The data in each data mart is usually tailored for a particular capability or function, such as product profitability analysis, KPI analyses, customer demographic analyses, and so on.
Technical metadata describes the physical structures in the CIF and the detailed processes that move and transform data in the environment.

Business metadata describes the data structures, data elements, business rules, and business usage of data in the CIF.

Administrative metadata describes the operation of the CIF, including audit trails, performance metrics, data quality metrics, and other statistical metadata.

Information feedback is the sharing mechanism that allows intelligence and knowledge gathered through the usage of the Corporate Information Factory to be shared with other data stores, as appropriate.
The library component provides a directory of the resources and data available in the CIF, organized in a way that makes sense to business users. This directory is much like a library, in that there is a standard taxonomy for categorizing and ordering information components.

**toolbox** is the collection of reusable components (for example, analytical reports) that business users can share, in order to leverage work and analysis performed by others in the enterprise.

In the **workbench**, metadata, data, and analysis tools are organized around business functions and tasks that supports business users in their jobs.

---

**Role and Purpose of the Data Warehouse**

[Imhoff, 2003]
The multipurpose nature of the DW

- It should be enterprise focused
- Its design should be as resilient to change as possible.
- It should be designed to load massive amounts of data in very short amounts of time.
- It should be designed for optimal data extraction processing by the data delivery programs.
- Its data should be in a format that supports any and all possible BI analyses in any and all technologies.

Design Pattern for the DW

- Non-redundant
- Stable
  - since change is inevitable, we must be prepared to accommodate newly discovered entities or attributes as new BI capabilities and data marts are created.
- Consistent
- Flexible in Terms of the Ultimate Data Usage
Standard ER approach + Historical Data + Structures Changes

Quick overview of OLAP cube concepts
A business sells **products** in **stores** and it is necessary to measure the company’s performance through **time**.

Values concerning
- a product
- a day
- a store

**Dollar Sales amount**
**Unit Sales**

**Time (days)**

**Products**

**Stores**

**Hyper-Cube**

**Sparce Matrix**
**Basic operation: Slice**

*Slice*: a subset of multidimensional data

*Slice*: a slice is defined by selecting specific values of dimension’s attributes

---

**Basic operation: Aggregation**

\[
\sum_{l=t_2, t=t_1} f(l, p, t)
\]

\[
\sum_{l=t_1, p \in \text{MarcaX}} f(l, p, t)
\]

\[
\sum_{l=t_1, p \in \text{MarcaY}} f(l, p, t)
\]

\[
\sum_{l \in \{t_2, t_3, t_5\}, t=t_1} f(l, p, t)
\]

Região 1
Basics of Multidimensional Modeling

Multidimensional Cube

- A Data Modeling approach with the purpose of addressing the following aspects:
  - The resulting data models should be understandable by the analytical users:
    - Simple.
    - Using terms from the domain and appropriate for data analysis.
  - Provides a framework for efficient querying
  - Provides the basics for generic software development where the users can navigate in large data sets in an intuitive way
Star schema

- **Fact table**
  - Big and central table. The only table with many joins connecting with the others tables.

- **Many Dimension Tables**
  - With only one join connecting to the fact table.

---

Fact Tables

- **Numerical measures of process.**
  - Continuos values (or represented as continuos values).
  - Additive (may be correctly added by any dimension).
  - Semi-additive (may be correctly added by some dimension but not on other dimensions).
  - Non-additive (cannot be added but some other aggregation operators are allowed).

- The goal is to summarize the information presented in fact tables.

- The granularity of a fact table is defined by a sub-set of dimensions that index it.
  - Ex: sales per day, store and product.

- Fact tables are, in general, sparse.
  - Ex: If a product is not sold on a day, in a store then there is no correspondent record on the fact table.
Dimension Tables

- Tables with simple primary keys that are related to fact tables.
- The most interesting attributes are the ones with textual descriptions.
  - They are used to define constraints over the data that will be analyzed.
  - They are used to group the aggregations made over the fact table measures. They will be the header’s columns.

### Brand | Dollar amount sold | Sold Units
--- | --- | ---
M-1 | 780 | 263
M-2 | 1044 | 509
M-3 | 213 | 444
M-4 | 95 | 39

Typical result

- Data for the first quarter for all stores by brand

### Brand | Dollar amount sold | Sold Units
--- | --- | ---
M-1 | 780 | 263
M-2 | 1044 | 509
M-3 | 213 | 444
M-4 | 95 | 39

Distinct values for the selected attribute

Textual Attribute of a Dimension
Querying a Star Schema

 DW Basic Concepts - 31

Basics of Multidimensional Modeling

Typical SQL query for OLAP

```
select p.brand, sum(f.value), sum(f.units)  
from sales f, product p, time t  
where f.product_key = p.product_key  
     and f.time_key = t.time_key  
     and f.quarter = "Q1 1996"  
group by p.brand  
order by p.brand
```

Selecting the columns

Join constraint

Application constraint

Grouping

Sorting
Processing the SQL query for OLAP

- First, the application constraints are processed for each dimension
  - Ex: Month = “Mars”; Year = 1997; Type of store = “Hyper”; Region = “..”; ...

- Each dimension produces a set of candidate keys:
  - Ex: Time: All time_key for which Month = “Mars”; Year = 1997;

- All the candidate keys are concatenated (Cartesian Product) to get the keys to be searched in the fact tables.

- All the hits on the fact table are grouped and aggregated.

Browsing the Dimension Tables

- “Dimension Browsing” - is the user activity where the user explore the data in the dimensions with the purpose of defining constraints over the dimension’s attributes and to select the level and type of intended summarization for the OLAP answers.

- Generic and convenient mechanism used by the user to specify the Queries.

  - SIMPLICITY
  - PERFORMANCE
Browsing the Dimension Tables

**Dimensions: dim1 (ex: produto)**

<table>
<thead>
<tr>
<th>Atributo:</th>
<th>Marca</th>
<th>Tipo</th>
<th>Nome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restrição:</td>
<td>Alcatel</td>
<td>Nokia</td>
<td>Telemóvel</td>
</tr>
<tr>
<td>Valores Distintos:</td>
<td>Alcatel</td>
<td>Ericson</td>
<td>Nokia</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>Telemóvel</td>
<td>Televisão</td>
</tr>
</tbody>
</table>

Drill Down e Drill Up

<table>
<thead>
<tr>
<th>Department</th>
<th>Sales Amount</th>
<th>Sales Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-1</td>
<td>780</td>
<td>263</td>
</tr>
<tr>
<td>D-2</td>
<td>1044</td>
<td>509</td>
</tr>
<tr>
<td>D-3</td>
<td>213</td>
<td>444</td>
</tr>
<tr>
<td>D-4</td>
<td>95</td>
<td>39</td>
</tr>
</tbody>
</table>

Drill down to department and Brand

<table>
<thead>
<tr>
<th>Department</th>
<th>Brand</th>
<th>Sales Amount</th>
<th>Sales Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-1</td>
<td>M-1</td>
<td>300</td>
<td>160</td>
</tr>
<tr>
<td>D-1</td>
<td>M-2</td>
<td>480</td>
<td>103</td>
</tr>
<tr>
<td>D-2</td>
<td>M-5</td>
<td>...</td>
<td>.....</td>
</tr>
<tr>
<td>...</td>
<td>......</td>
<td>...</td>
<td>.....</td>
</tr>
</tbody>
</table>
Drill Down e Drill Up

- Drill down is just to add some new header columns to the result table, which is a dimension attribute.

- Drill-Up is the reverse operations.

---

From a rowset to an analytical view

![Diagram showing dimensional attributes and variables](image-url)
**Classical OLAP view**

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th></th>
<th>Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Toys</td>
<td>Clothes</td>
<td>Toys</td>
</tr>
<tr>
<td>Sales</td>
<td>Costs</td>
<td>Sales</td>
<td>Costs</td>
</tr>
<tr>
<td>Q1</td>
<td>320</td>
<td>200</td>
<td>825</td>
</tr>
<tr>
<td>Q2</td>
<td>225</td>
<td>220</td>
<td>390</td>
</tr>
<tr>
<td>Q3</td>
<td>700</td>
<td>600</td>
<td>425</td>
</tr>
<tr>
<td>Q4</td>
<td>880</td>
<td>850</td>
<td>875</td>
</tr>
</tbody>
</table>

**Inefficient OLAP view**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Actual</th>
<th></th>
<th>Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Toys</td>
<td></td>
<td>Clothes</td>
<td></td>
</tr>
<tr>
<td>Sales</td>
<td>Costs</td>
<td>Sales</td>
<td>Costs</td>
<td>Sales</td>
</tr>
<tr>
<td>Q1</td>
<td>320</td>
<td>205</td>
<td>700</td>
<td>600</td>
</tr>
<tr>
<td>Q2</td>
<td>225</td>
<td>200</td>
<td>630</td>
<td>554</td>
</tr>
<tr>
<td>Q3</td>
<td>700</td>
<td>625</td>
<td>875</td>
<td>653</td>
</tr>
<tr>
<td>Q4</td>
<td>880</td>
<td>850</td>
<td>890</td>
<td>893</td>
</tr>
</tbody>
</table>

**Basics of Multidimensional Modeling**

*DW Basic Concepts - 39*

*DW Basic Concepts - 40*
<table>
<thead>
<tr>
<th>Division</th>
<th>Tipo</th>
<th>PROD</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESS</td>
<td>Audio</td>
<td>Tape</td>
<td>1218213</td>
<td>1193207</td>
<td>70166</td>
<td>835466</td>
<td>39430106</td>
</tr>
<tr>
<td></td>
<td>Audio</td>
<td>CAS</td>
<td>131739</td>
<td>125868</td>
<td>688</td>
<td>93642</td>
<td>419449</td>
</tr>
<tr>
<td></td>
<td>Metal</td>
<td>CAS</td>
<td>83354</td>
<td>825847</td>
<td>48366</td>
<td>550266</td>
<td>2693327</td>
</tr>
<tr>
<td></td>
<td>Stn</td>
<td>CAS</td>
<td>248119</td>
<td>241493</td>
<td>14916</td>
<td>191558</td>
<td>8033</td>
</tr>
<tr>
<td>Accessory</td>
<td>TV</td>
<td>Total</td>
<td>242545</td>
<td>235641</td>
<td>14024</td>
<td>167092</td>
<td>7888212</td>
</tr>
<tr>
<td>Accessory</td>
<td>VHS/Video</td>
<td>Total</td>
<td>965751</td>
<td>102288</td>
<td>5473</td>
<td>63926</td>
<td>3171037</td>
</tr>
<tr>
<td>Accessory</td>
<td>DVD</td>
<td>Video</td>
<td>1073928</td>
<td>106047</td>
<td>57785</td>
<td>714094</td>
<td>3425919</td>
</tr>
<tr>
<td>Accessory</td>
<td>Camcorder</td>
<td>Video</td>
<td>639691</td>
<td>647732</td>
<td>40578</td>
<td>539556</td>
<td>225222</td>
</tr>
<tr>
<td>Accessory</td>
<td>TV</td>
<td>Total</td>
<td>267937</td>
<td>272926</td>
<td>152736</td>
<td>1912818</td>
<td>8849176</td>
</tr>
</tbody>
</table>

**SUM OF SALES**

<table>
<thead>
<tr>
<th>Division</th>
<th>Tipo</th>
<th>PROD</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessory</td>
<td>TV</td>
<td>Total</td>
<td>310499</td>
<td>311604</td>
<td>29306</td>
<td>58875</td>
<td>16735388</td>
</tr>
<tr>
<td>Audio</td>
<td>Audio</td>
<td>Comp</td>
<td>1088763</td>
<td>5971602</td>
<td>54424</td>
<td>624428</td>
<td>23333699</td>
</tr>
<tr>
<td></td>
<td>Amplifier</td>
<td>Comp</td>
<td>2043401</td>
<td>7176682</td>
<td>85518</td>
<td>1136034</td>
<td>5750897</td>
</tr>
<tr>
<td></td>
<td>CD Player</td>
<td>Comp</td>
<td>1483033</td>
<td>1214974</td>
<td>597363</td>
<td>7890646</td>
<td>4084593</td>
</tr>
<tr>
<td></td>
<td>Receiver</td>
<td>Comp</td>
<td>6498612</td>
<td>9089041</td>
<td>507632</td>
<td>6006966</td>
<td>2881886</td>
</tr>
<tr>
<td></td>
<td>Tuner</td>
<td>Comp</td>
<td>288388</td>
<td>261363</td>
<td>137244</td>
<td>1675234</td>
<td>8549498</td>
</tr>
<tr>
<td>Audio</td>
<td>Comp</td>
<td>Total</td>
<td>3929107</td>
<td>3954605</td>
<td>187035</td>
<td>22051866</td>
<td>116492772</td>
</tr>
<tr>
<td>Portaudio</td>
<td>Audio</td>
<td>PortCAS</td>
<td>216937</td>
<td>2239396</td>
<td>117208</td>
<td>1638688</td>
<td>7293071</td>
</tr>
<tr>
<td></td>
<td>PortCD</td>
<td>Audio</td>
<td>371393</td>
<td>3016612</td>
<td>138033</td>
<td>1800258</td>
<td>9991163</td>
</tr>
<tr>
<td></td>
<td>PortST</td>
<td>Audio</td>
<td>3024177</td>
<td>318752</td>
<td>174462</td>
<td>21478</td>
<td>10103749</td>
</tr>
<tr>
<td>Portaudio</td>
<td>Audio</td>
<td>Total</td>
<td>8923867</td>
<td>8494746</td>
<td>429703</td>
<td>5586926</td>
<td>27305283</td>
</tr>
<tr>
<td>Audio</td>
<td>TV</td>
<td>Total</td>
<td>4621495</td>
<td>4403475</td>
<td>2350059</td>
<td>2838782</td>
<td>14379808</td>
</tr>
<tr>
<td>Video</td>
<td>Camorder</td>
<td>B3</td>
<td>1277086</td>
<td>1220167</td>
<td>680154</td>
<td>827122</td>
<td>5979523</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>NMC</td>
<td>9030893</td>
<td>934343</td>
<td>452323</td>
<td>563122</td>
<td>2853076</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>VSNCC</td>
<td>154074</td>
<td>1472182</td>
<td>815937</td>
<td>977732</td>
<td>4806634</td>
</tr>
<tr>
<td>Camorder</td>
<td>Total</td>
<td>3720971</td>
<td>3626687</td>
<td>1928394</td>
<td>2363724</td>
<td>11639257</td>
<td></td>
</tr>
<tr>
<td>TV</td>
<td>B1</td>
<td>BW TV</td>
<td>114263</td>
<td>1159484</td>
<td>6875</td>
<td>851242</td>
<td>3859856</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>Color TV</td>
<td>23693</td>
<td>198461</td>
<td>10171</td>
<td>12954</td>
<td>6601179</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>PortTV</td>
<td>159184</td>
<td>1451</td>
<td>72659</td>
<td>786424</td>
<td>4555965</td>
</tr>
<tr>
<td>TV</td>
<td>Total</td>
<td>510349</td>
<td>463429</td>
<td>240587</td>
<td>2933718</td>
<td>1507677</td>
<td></td>
</tr>
<tr>
<td>VCR</td>
<td>B2</td>
<td>STN VCR</td>
<td>211997</td>
<td>198156</td>
<td>119101</td>
<td>135693</td>
<td>664955</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>STR VCR</td>
<td>3781857</td>
<td>390457</td>
<td>190967</td>
<td>230196</td>
<td>11897693</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>TOTALProd</td>
<td>0928324</td>
<td>575747</td>
<td>1258883</td>
<td>4046701</td>
<td>19013895</td>
</tr>
<tr>
<td>VCR</td>
<td>Total</td>
<td>6548012</td>
<td>6346102</td>
<td>3566891</td>
<td>4412535</td>
<td>208662</td>
<td></td>
</tr>
<tr>
<td>Video</td>
<td>TV</td>
<td>Total</td>
<td>10774283</td>
<td>10436218</td>
<td>5735932</td>
<td>7069048</td>
<td>340155267</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>TV</td>
<td>14106274</td>
<td>13532799</td>
<td>8329043</td>
<td>10281249</td>
<td>5008487</td>
</tr>
</tbody>
</table>

**DATA WAREHOUSE - BASIC CONCEPTS**

**BASICS OF MULTIDIMENSIONAL MODELING**

**FURTHER READING AND SUMMARY**
Further Reading and Summary

Readings

- (Kimball - The Data Warehouse toolkit, 2002) - pag 16 to 27.

What you should know:

- Understand the Corporate Information Model (CIF): The different roles for the DW, the ODS and the Data Marts (specially the OLAP data marts). The fundamental aspect of feedback from the knowledge and information gathered at DSS systems into the architecture (operational systems and the DW).

- Understand the fundamental differences between OLTP and the analytical activities developed on the DW or on the Data Marts: data, access, users ...