02

Data Warehouse - Basic Concepts

Notice

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

Bibliography

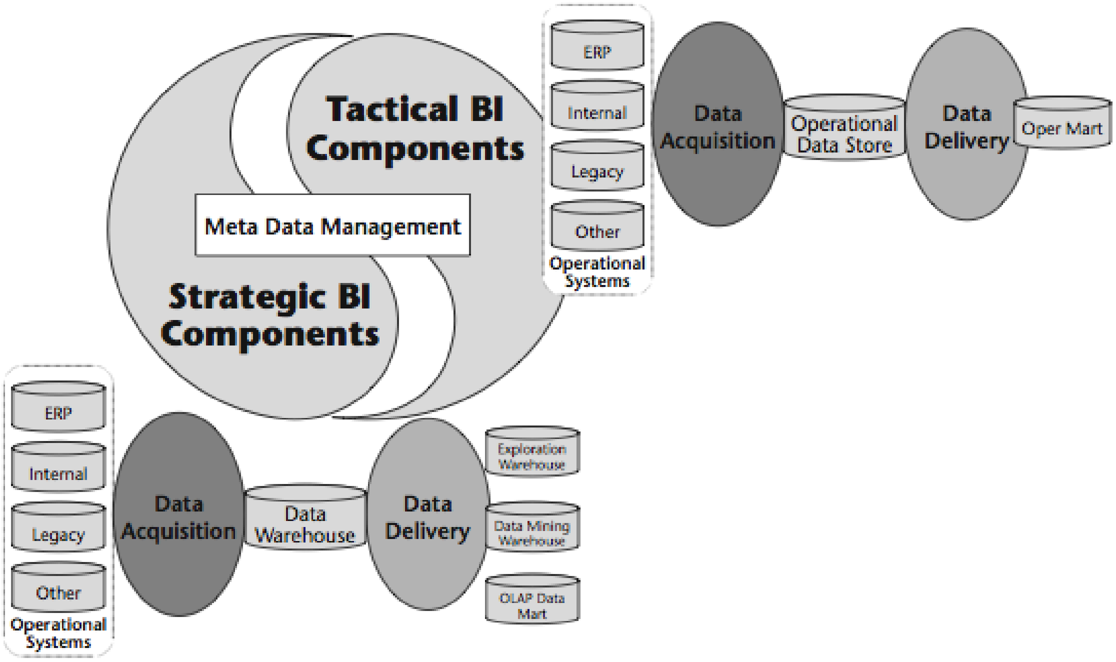
- Many examples are extracted and adapted from
  - [Imhoff , 2003] - Mastering Data Warehouse Design : Relational and Dimensional Techniques, Wiley.
  - [Kimball, 2002] - The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition), from Ralph Kimball, Margy Ross, Willey

Table of Contents

- Corporate Information Factory
- Quick overview of OLAP cube concepts
- Basics of Multidimensional Modeling

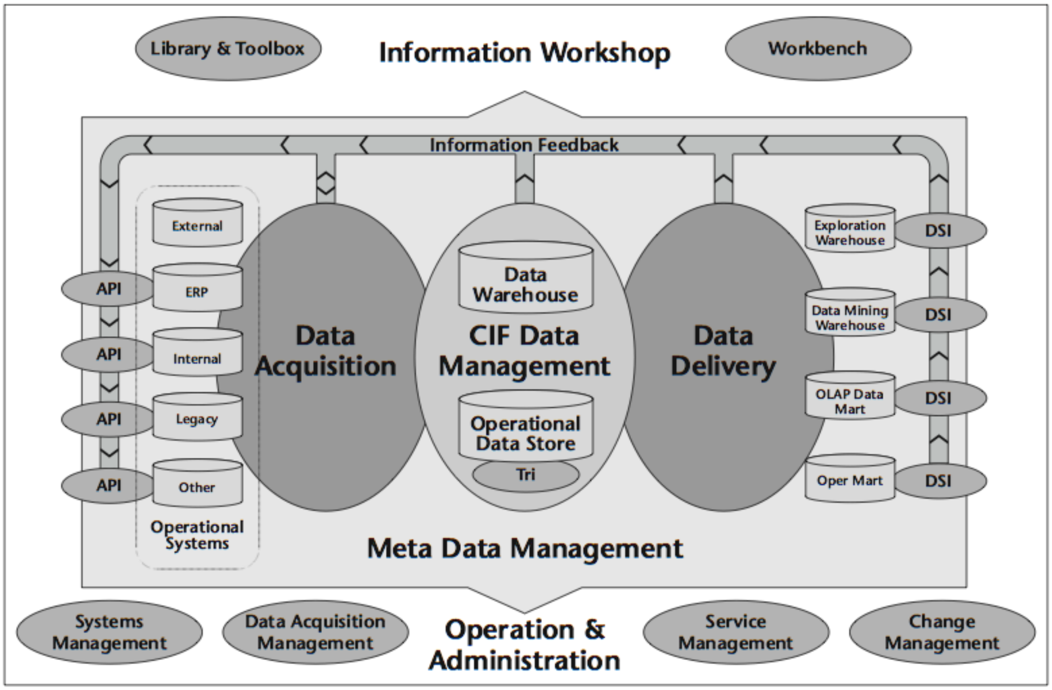
Corporate Information Factory

Strategic and tactical portions of a BI environment.



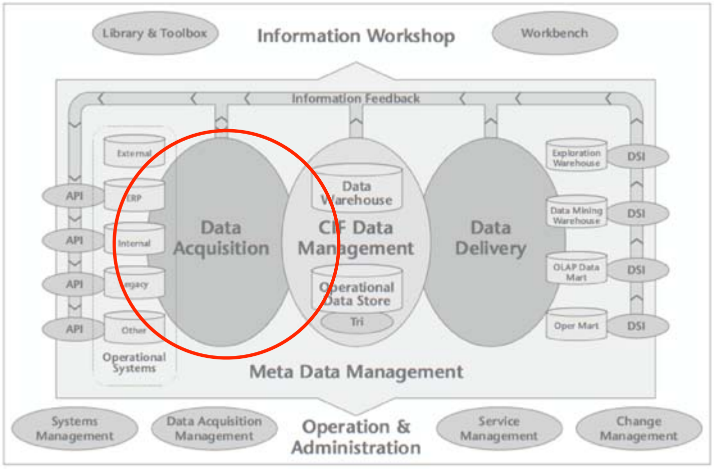
[Imhoff , 2003]

Corporate Information Factory Architecture



[Imhoff , 2003]

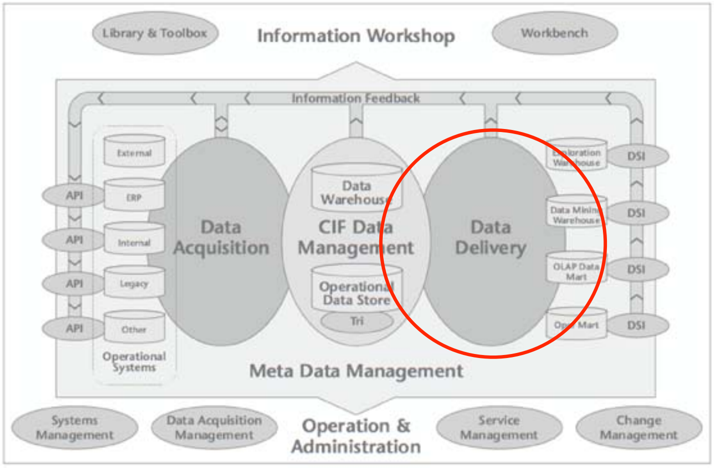
CIF: Data Acquisition - (ETL)



[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.

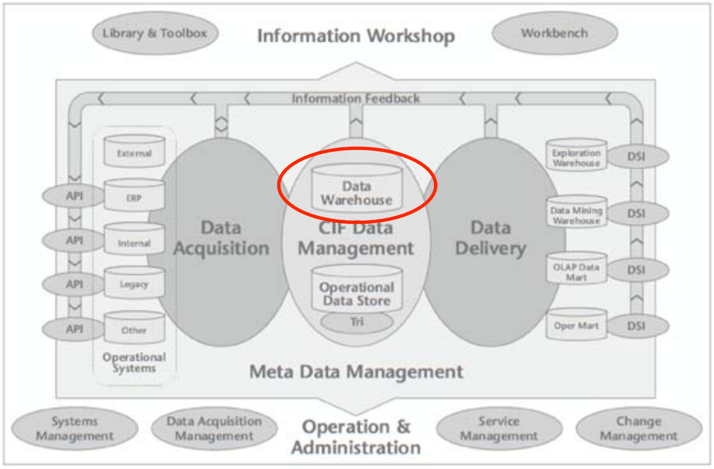
CIF: Data Delivery - (ETL)



[Imhoff , 2003]

Data delivery is the process that moves data from the data warehouse into data and oper 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 highquality, integrated data that conforms to the enterprise business rules.

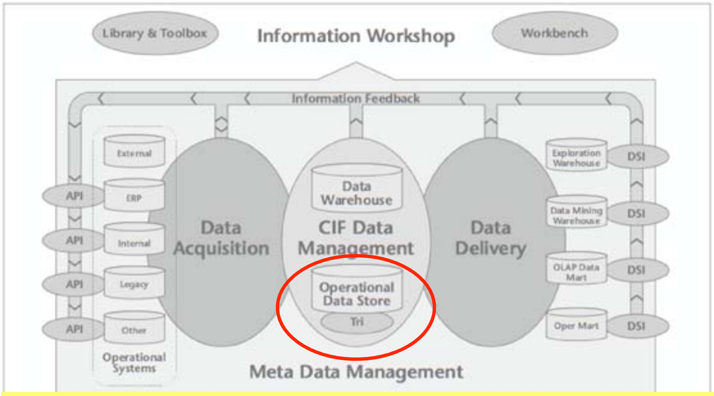
CIF: Data Warehouse



[Imhoff , 2003]

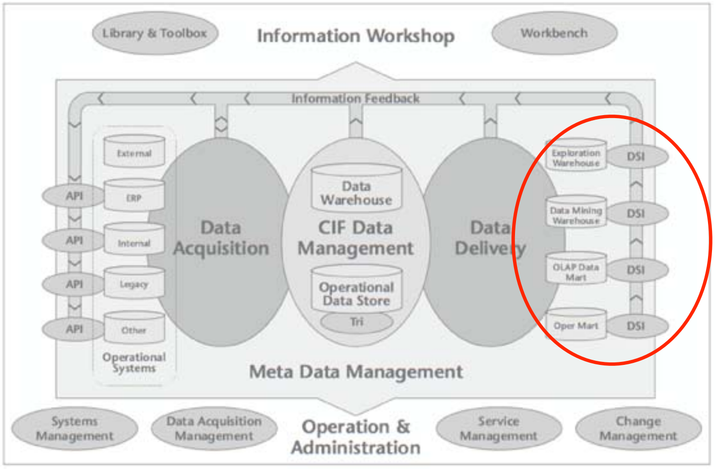
“a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making” [Imnon, 1980]

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

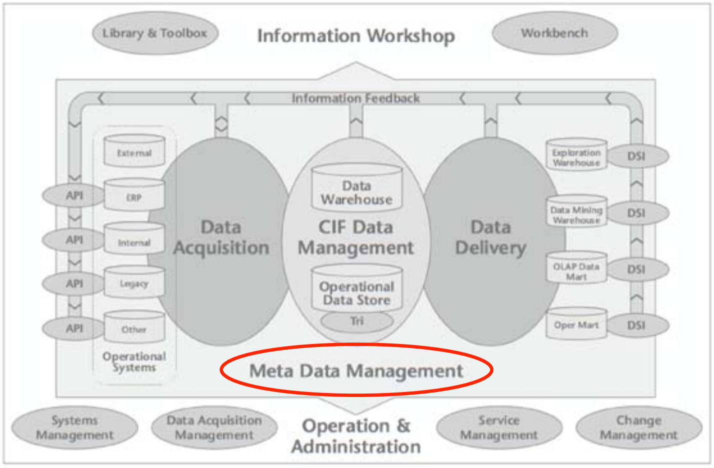
CIF: Data Mart



[Imhoff , 2003]

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.

CIF: Metadata Management



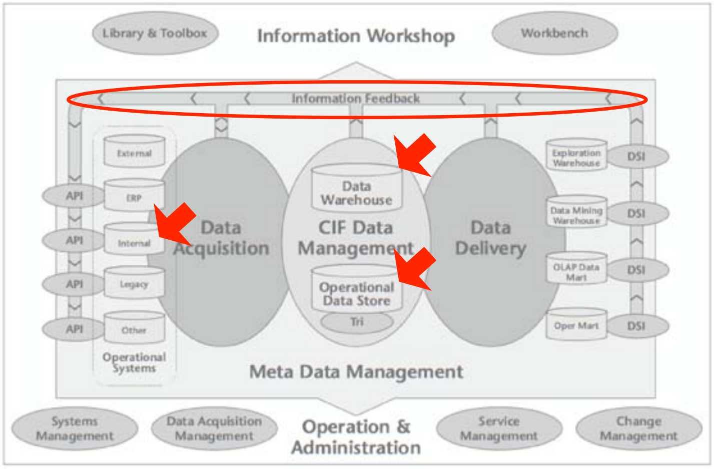
[Imhoff , 2003]

Technical meta data 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 meta data.

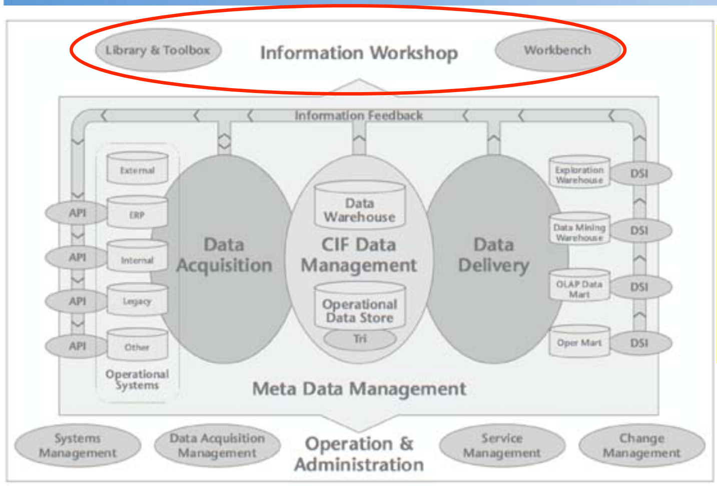
CIF: Information feedback



[Imhoff , 2003]

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

CIF: Information Workshop



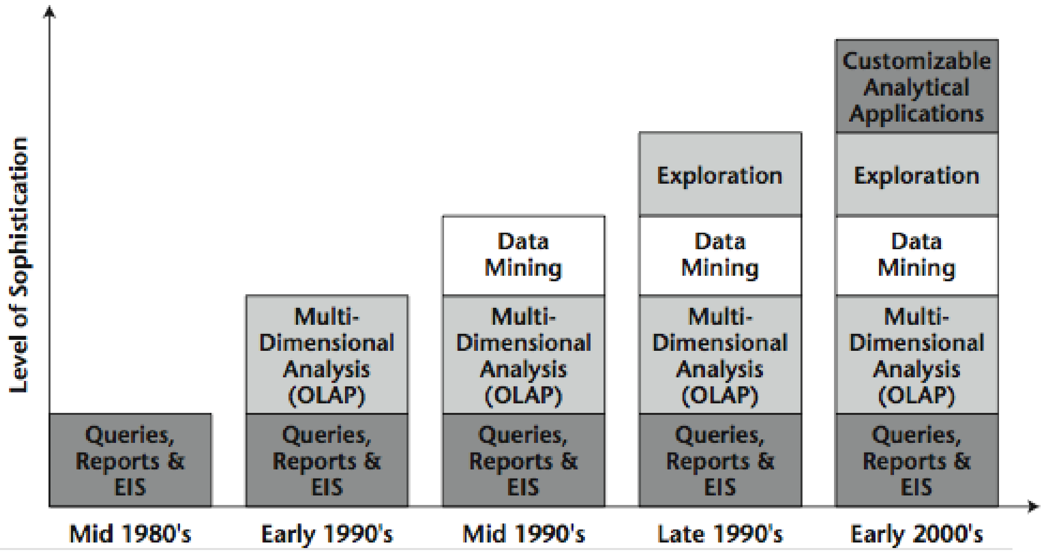
The **library component** provides a directory of the resources and data available in the CIF, organized in a way that makes sense to business users. This directory is much like a library, in that there is a standard taxonomy for categorizing and ordering information components.

[Imhoff , 2003]

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**

Design Pattern for the DW

Standard ER approach

+

Historical Data

+

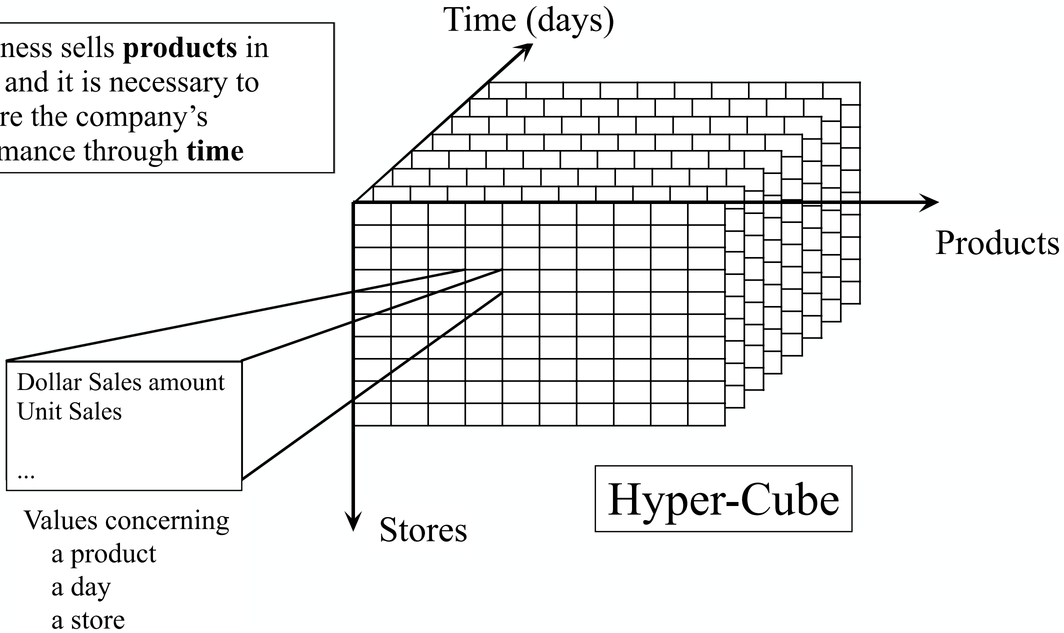
Structures Changes

Data Warehouse - Basic Concepts

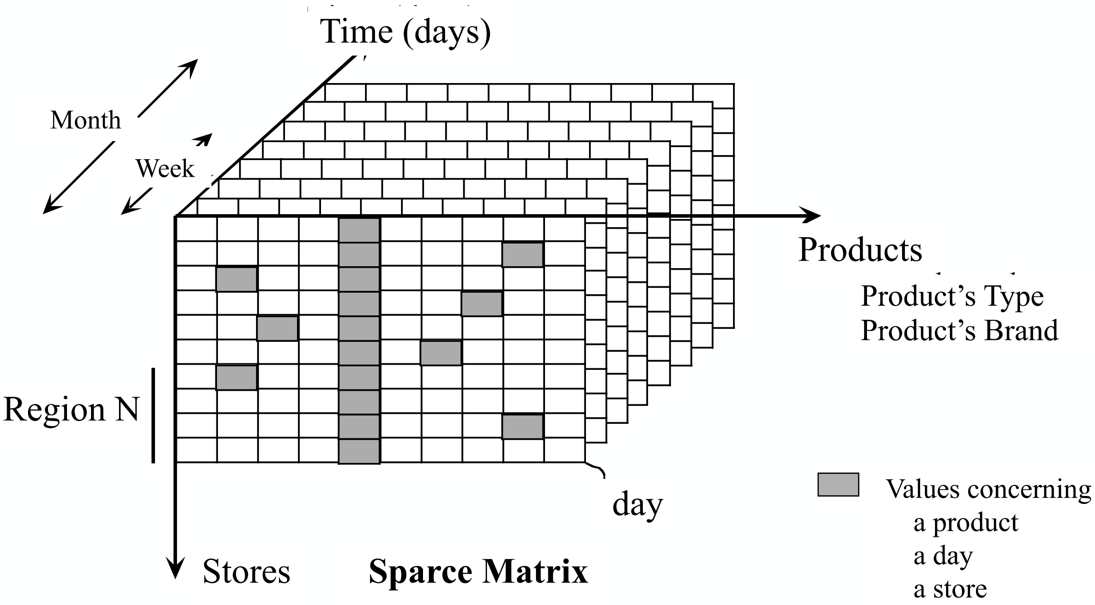
Quick overview of OLAP cube concepts

Multidimensional Cube

A business sells **products** in **stores** and it is necessary to measure the company's performance through **time**



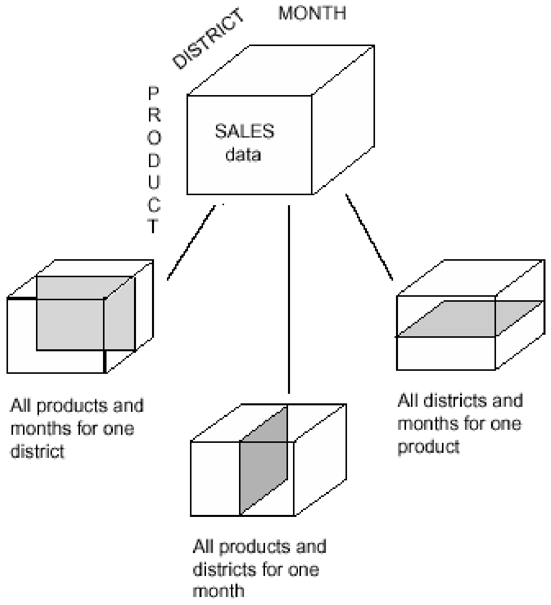
Multidimensional Cube



Basic operation: Slice

Slice: a subset of multidimensional data

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

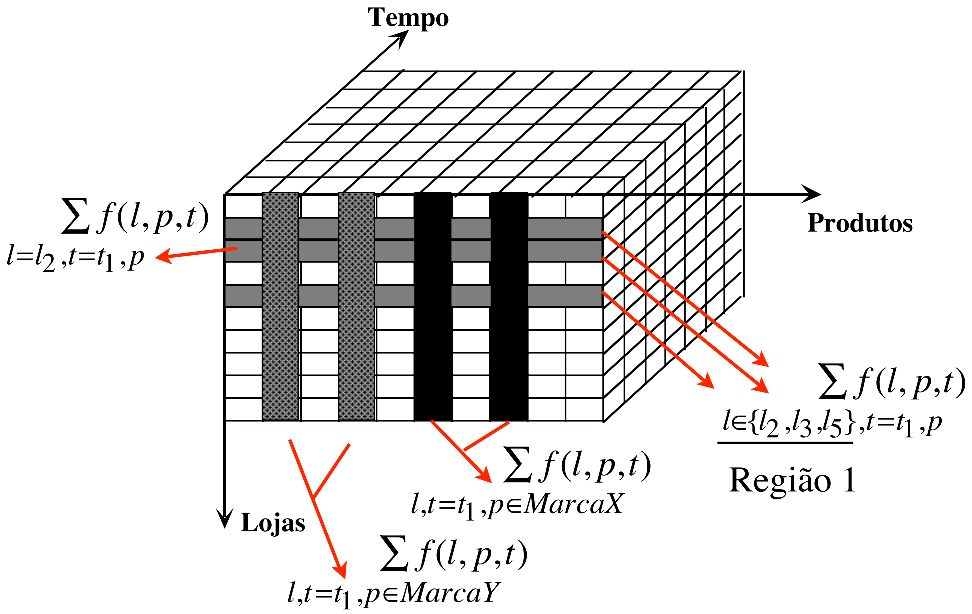


All products and months for one district

All products and districts for one month

All districts and months for one product

Basic operation: Aggregation


Tempo
Produtos
Lojas
$\sum_{l=l_2,t=t_1,p} f(l,p,t)$
$\sum_{l\in\{l_2,l_3,l_5\},t=t_1,p} f(l,p,t)$
Região 1
$\sum_{l,t=t_1,p\in MarcaX} f(l,p,t)$
$\sum_{l,t=t_1,p\in MarcaY} f(l,p,t)$

Basics of Multidimensional Modeling

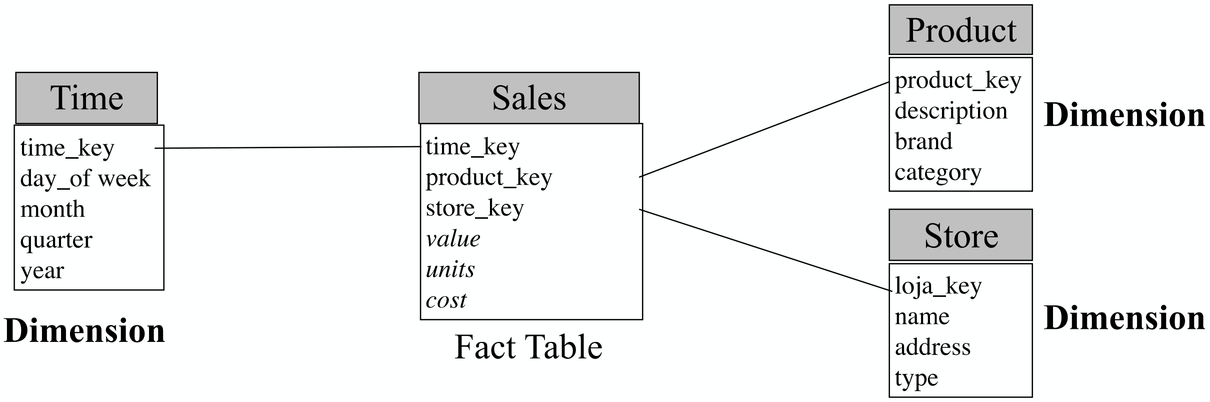
Multidimensional Cube

- **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**

Asymmetric Model



Time

time_key
day_of week
month
quarter
year

Dimension

Sales

time_key
product_key
store_key
value
units
cost

Fact Table

Product

product_key
description
brand
category

Dimension

Store

loja_key
name
address
type

Dimension

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.**

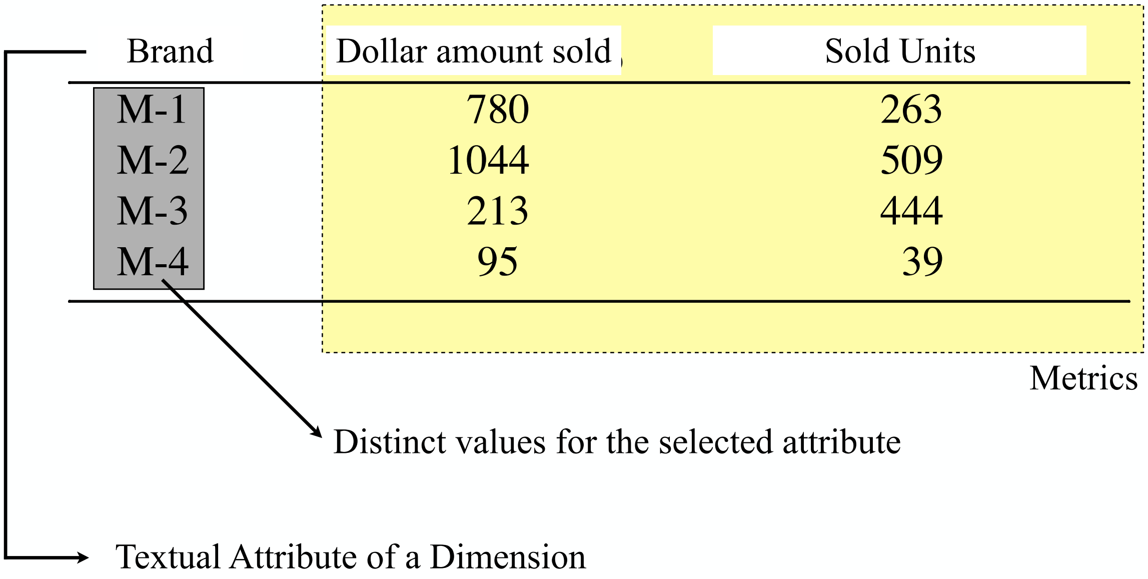
Dimension Tables

- **Tables with simple primary keys that are related to fact tables.**
- **The most interesting attributes 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**



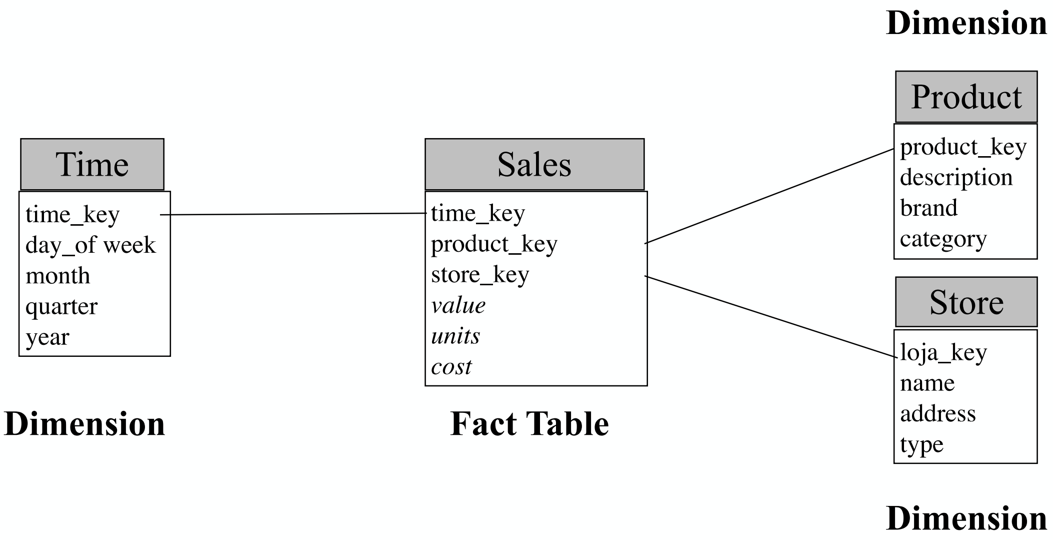
Brand	Dollar amount sold	Sold Units
M-1	780	263
M-2	1044	509
M-3	213	444
M-4	95	39

Metrics

Distinct values for the selected attribute

Textual Attribute of a Dimension

Querying a Star Schema



Typical SQL query for OLAP

Selecting the columns

```
select p.brand, sum(f.value), sum(f.units)
from sales f, product p, time t                ←  aliases

where f.product_key = p.product_key            ←  Join constraint
      and f.time_key = t.time_key              ←  Join constraint
      and f.quarter = “Q1 1996”                ←  Application constraint

group by p.brand                               ←  Grouping
order by p.brand                               ←  Sorting
```

Processing the SQL query for OLAP

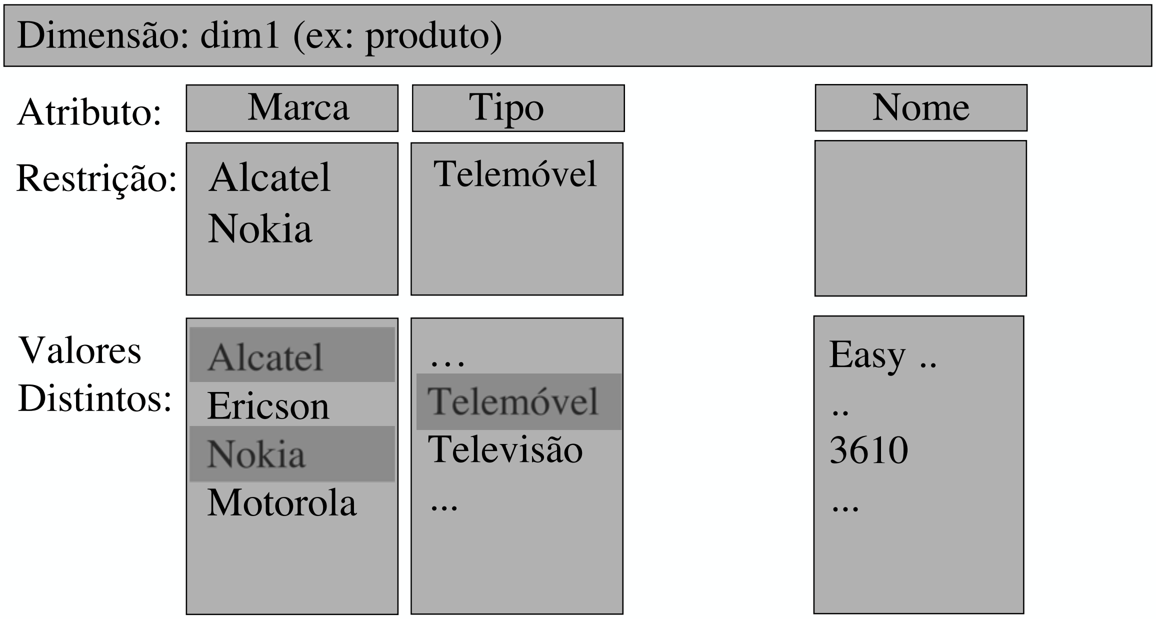
- First, the application constraints are processed for each dimension
  - Ex: Month = “Mars”; Year = 1997; Type of store = “Hyper”;

    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



Dimensão: dim1 (ex: produto)

Atributo:	Marca	Tipo	Nome
Restrição:	Alcatel Nokia	Telemóvel	
Valores Distintos:	Alcatel Ericson Nokia Motorola	… Telemóvel Televisão ...	Easy .. .. 3610 ...

Drill Down e Drill Up

Department	Sales Amount	Sales Units
D-1	780	263
D-2	1044	509
D-3	213	444
D-4	95	39

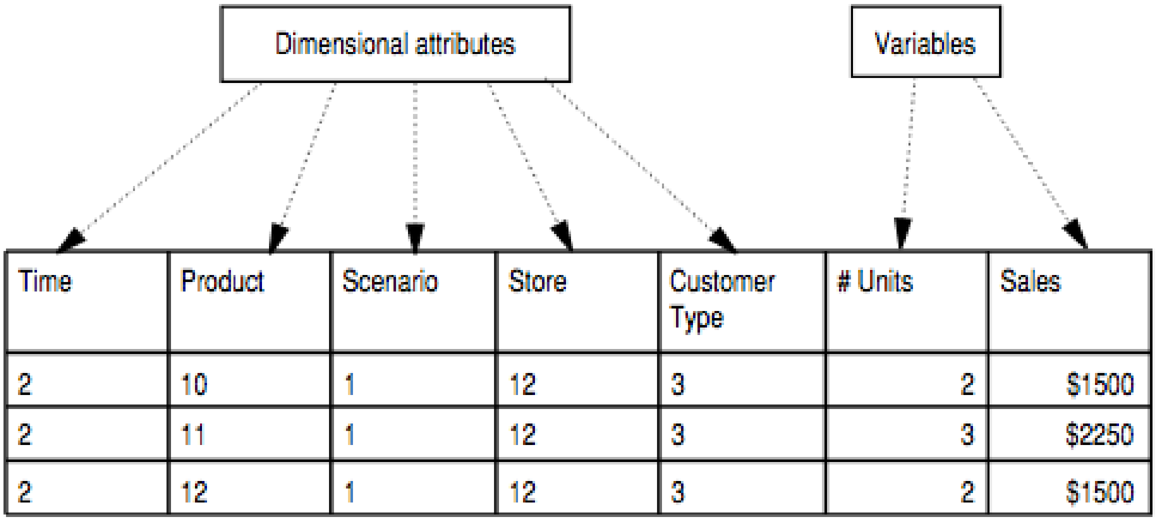
Drill down to department and Brand

Department	Brand	Sales Amount	Sales Units
D-1	M-1	300	160
D-1	M-2	480	103
D-2	M-5	...	…..
...	……	….	…..

Drill Down e Drill Up

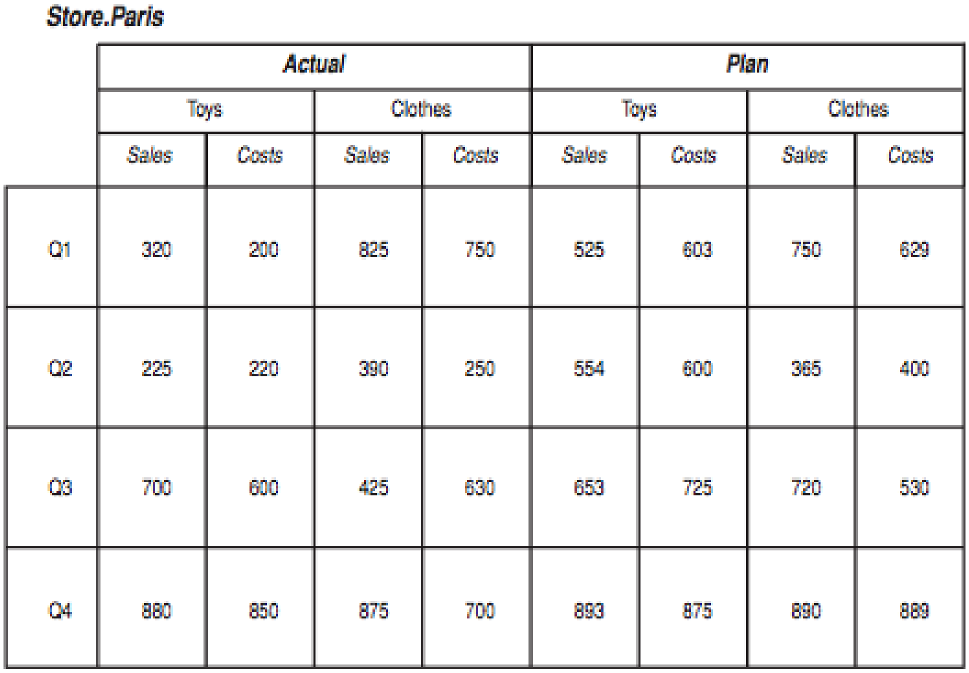
- **Drill down is just to add some new header columns to the result table, which is a dimension attribute**
- **Drill-Up is the reverse operations**

From a rowset to an analytical view



Time	Product	Scenario	Store	Customer Type	# Units	Sales
2	10	1	12	3	2	$1500
2	11	1	12	3	3	$2250
2	12	1	12	3	2	$1500

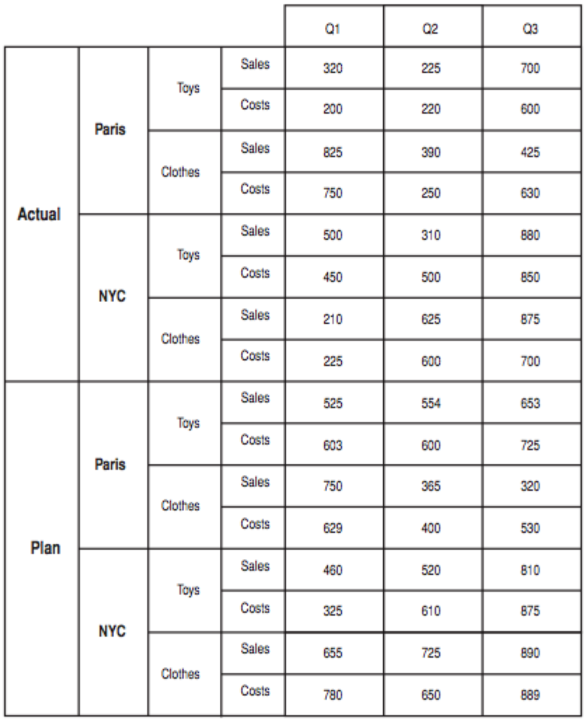
Classical OLAP view



Store.Paris

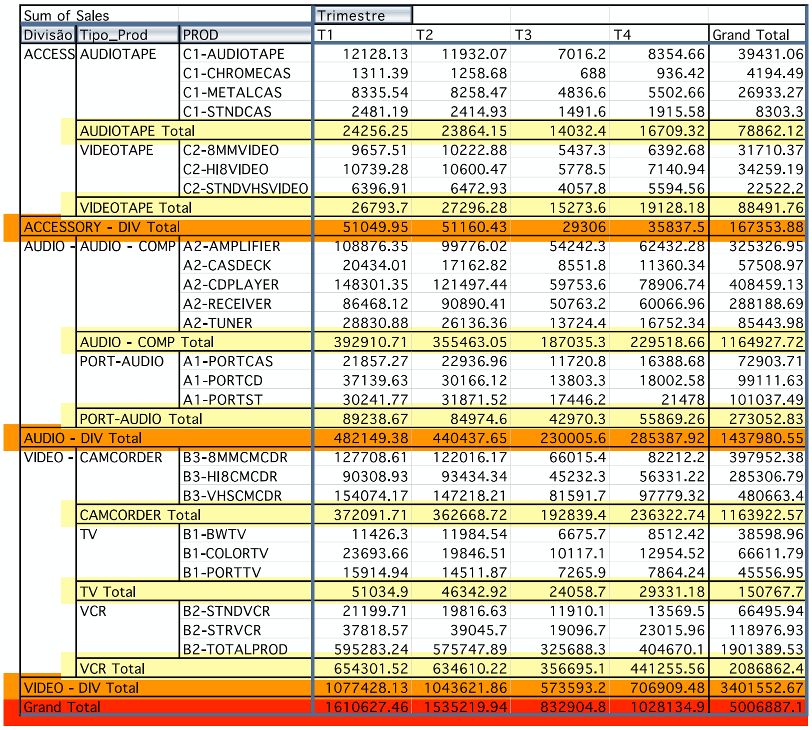
	Actual				*Plan*			
	Toys		Clothes		Toys		Clothes	
	Sales	*Costs*	*Sales*	*Costs*	*Sales*	*Costs*	*Sales*	*Costs*
Q1	320	200	825	750	525	603	750	629
Q2	225	220	390	250	554	600	365	400
Q3	700	600	425	630	653	725	720	530
Q4	880	850	875	700	893	875	890	889

Inefficient OLAP view



				Q1	Q2	Q3
Actual	**Paris**	Toys	Sales	320	225	700
			Costs	200	220	600
		Clothes	Sales	825	390	425
			Costs	750	250	630
	NYC	Toys	Sales	500	310	880
			Costs	450	500	850
		Clothes	Sales	210	625	875
			Costs	225	600	700
Plan	**Paris**	Toys	Sales	525	554	653
			Costs	603	600	725
		Clothes	Sales	750	365	320
			Costs	629	400	530
	NYC	Toys	Sales	460	520	810
			Costs	325	610	875
		Clothes	Sales	655	725	890
			Costs	780	650	889

What about Partial Totals?



Sum of Sales			Trimestre				
Divisão	Tipo_Prod	PROD	T1	T2	T3	T4	Grand Total
ACCESS	AUDIOTAPE	C1-AUDIOTAPE	12128.13	11932.07	7016.2	8354.66	39431.06
		C1-CHROMECAS	1311.39	1258.68	688	936.42	4194.49
		C1-METALCAS	8335.54	8258.47	4836.6	5502.66	26933.27
		C1-STNDCAS	2481.19	2414.93	1491.6	1915.58	8303.3
	AUDIOTAPE Total		24256.25	23864.15	14032.4	16709.32	78862.12
	VIDEOTAPE	C2-8MMVIDEO	9657.51	10222.88	5437.3	6392.68	31710.37
		C2-HI8VIDEO	10739.28	10600.47	5778.5	7140.94	34259.19
		C2-STNDVHSVIDEO	6396.91	6472.93	4057.8	5594.56	22522.2
	VIDEOTAPE Total		26793.7	27296.28	15273.6	19128.18	88491.76
ACCESSORY - DIV Total			51049.95	51160.43	29306	35837.5	167353.88
AUDIO -	AUDIO - COMP	A2-AMPLIFIER	108876.35	99776.02	54242.3	62432.28	325326.95
		A2-CASDECK	20434.01	17162.82	8551.8	11360.34	57508.97
		A2-CDPLAYER	148301.35	121497.44	59753.6	78906.74	408459.13
		A2-RECEIVER	86468.12	90890.41	50763.2	60066.96	288188.69
		A2-TUNER	28830.88	26136.36	13724.4	16752.34	85443.98
	AUDIO - COMP Total		392910.71	355463.05	187035.3	229518.66	1164927.72
	PORT-AUDIO	A1-PORTCAS	21857.27	22936.96	11720.8	16388.68	72903.71
		A1-PORTCD	37139.63	30166.12	13803.3	18002.58	99111.63
		A1-PORTST	30241.77	31871.52	17446.2	21478	101037.49
	PORT-AUDIO Total		89238.67	84974.6	42970.3	55869.26	273052.83
AUDIO - DIV Total			482149.38	440437.65	230005.6	285387.92	1437980.55
VIDEO -	CAMCORDER	B3-8MMCMCDR	127708.61	122016.17	66015.4	82212.2	397952.38
		B3-HI8CMCDR	90308.93	93434.34	45232.3	56331.22	285306.79
		B3-VHSCMCDR	154074.17	147218.21	81591.7	97779.32	480663.4
	CAMCORDER Total		372091.71	362668.72	192839.4	236322.74	1163922.57
	TV	B1-BWTV	11426.3	11984.54	6675.7	8512.42	38598.96
		B1-COLORTV	23693.66	19846.51	10117.1	12954.52	66611.79
		B1-PORTTV	15914.94	14511.87	7265.9	7864.24	45556.95
	TV Total		51034.9	46342.92	24058.7	29331.18	150767.7
	VCR	B2-STNDVCR	21199.71	19816.63	11910.1	13569.5	66495.94
		B2-STRVCR	37818.57	39045.7	19096.7	23015.96	118976.93
		B2-TOTALPROD	595283.24	575747.89	325688.3	404670.1	1901389.53
	VCR Total		654301.52	634610.22	356695.1	441255.56	2086862.4
VIDEO - DIV Total			1077428.13	1043621.86	573593.2	706909.48	3401552.67
Grand Total			1610627.46	1535219.94	832904.8	1028134.9	5006887.1

Data Warehouse - Basic Concepts

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 ...