01

Data Warehouse and OLAP

Notice

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

- **This material can be freely used for personal or academic purposes without any previous authorization from the author, provided that this notice is maintained/kept.**

- **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:**
  - **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling** (Third Edition). Ralph Kimball, Margy Ross.Wiley, 2013, ISBN: 978-1-118-53080-1
  - **Building the Data Warehouse** (4rd Edition), W. H. Inmon, Wiley, (4rd Edition), 2005, ISBN: 0-7645-9944-5
  - **Mastering Data Warehouse Design : Relational and Dimensional Techniques**. Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geiger. Wiley, 2003. ISBN: 0471324213

Table of Contents

Decision Support Systems

What is about this part of the course?

- **Data Warehouse (DW)**
  - **For now, lets think about DW as the Warehouse where all the important data is integrated and stored, including historical data, for future support of Data Analysis and Decision making**

- **On-line analytical processing (OLAP)**
  - **It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. The data comes from de DW**

[Wikipedia - DSS]

Decisions in the context of an organization?

- **Strategic decisions (long term)**
  - **Examples**
    - Analyzing the actual pattern buying to develop a new product;
    - Deciding the creation of a new university course.
- **Short term decisions - tactical decisions**
  - **Examples**
    - Changing the volumes of components to buy to our suppliers;
    - Analyzing the factors affecting the unsuccessful results of so many students.

Some **analysis patterns** used by OLAP users

- **Summarizing and aggregation of large amount of data**
- **Filtering, sorting, ranking**
- **Comparisons of different sets of data**
- **Search for outliers**
- **Analysis and discovery of patterns**
- **Analysis of tendencies in the data**

Who are the DW and OLAP users?

- **DSS analyst is a business person first and foremost, and a technician second.**
  - The primary job of the DSS analyst is to **define and discover information** used in corporate decision-making.
- **The DSS analyst has a mindset of "Give me what I say I want, then I can tell you what I really want.". In other words, the DSS analyst operates in a mode of discovery.**
  - this has a profound effect on the way the data warehouse is developed and on how systems using the data warehouse are developed.

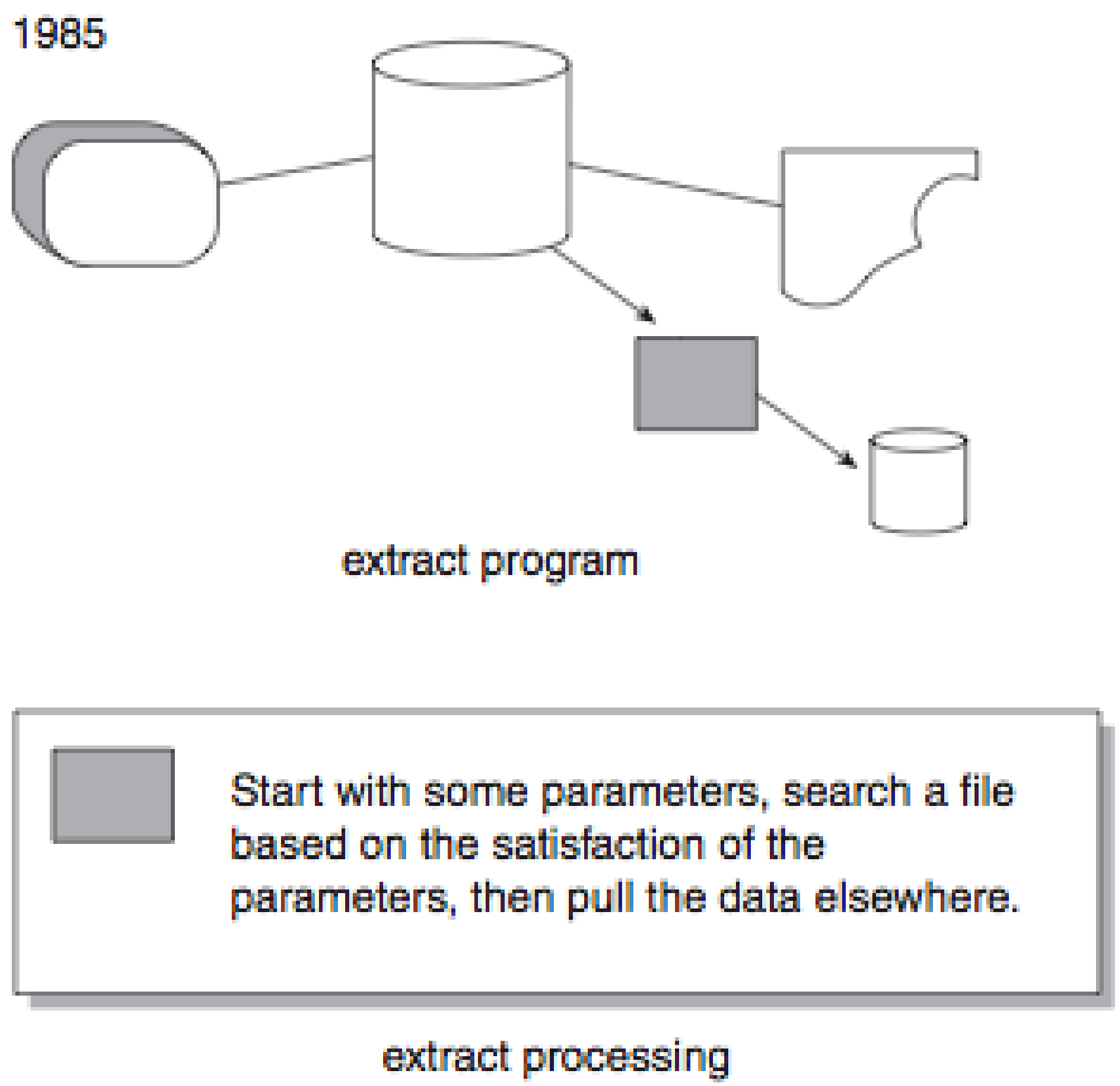
Historical perspective

Two different needs

- **Running the organization**
  - **Operational Data**
  - **Transactional Data**

- **Analyzing the organization performance**
  - **Aggregating Data**
  - **Comparing Data**

The “Extract” Program



[Inmon,2002]

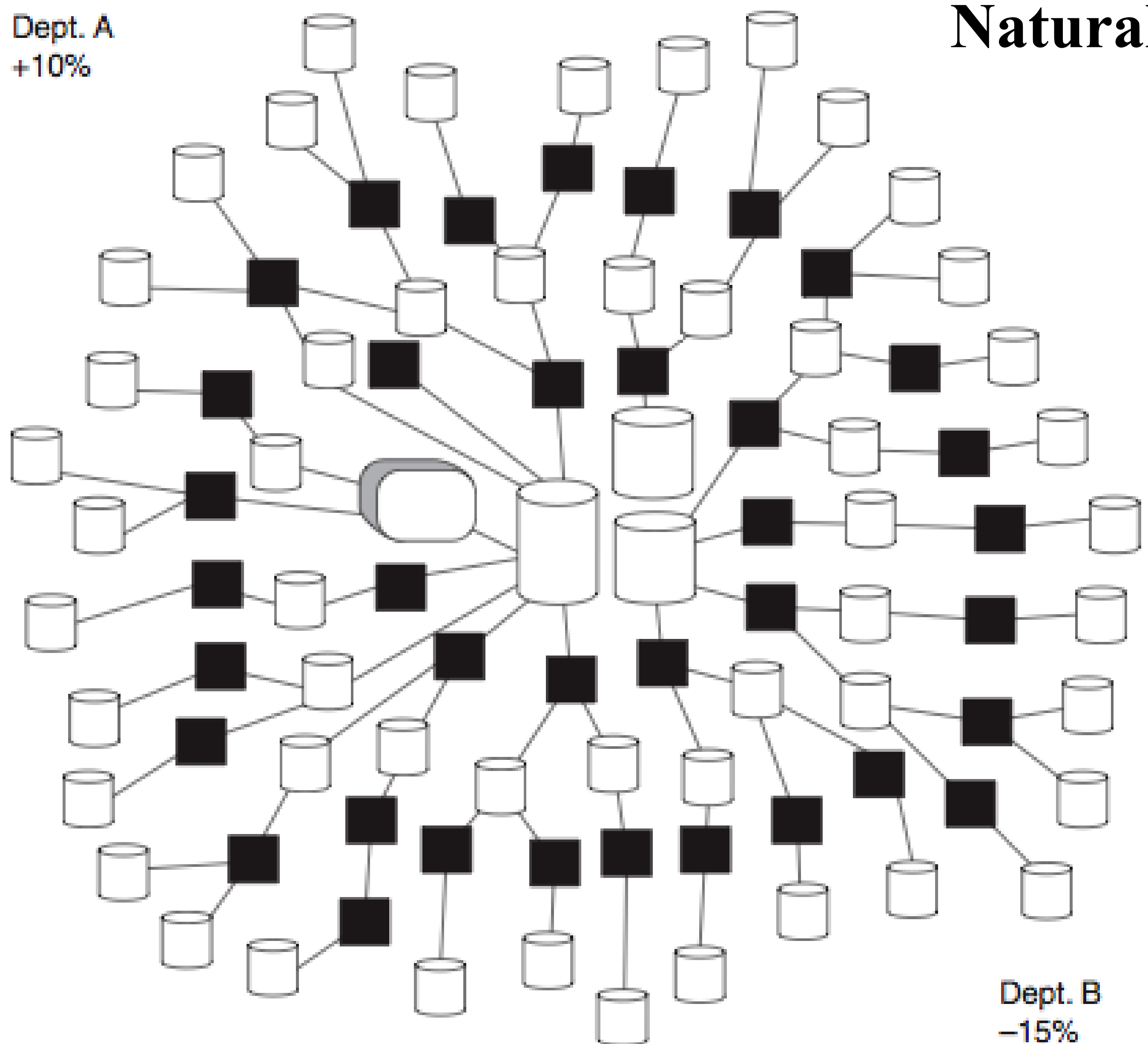
The "Extract" Program became very popular

- **Because extract processing can move data out of the way of high-performance online processing, there is no conflict in terms of performance when the data needs to be analyzed en masse.**

- **When data is moved out of the operational, transaction-processing domain with an extract program, a shift in control of the data occurs. The end user then owns the data once he or she takes control of it.**

The “Extract” Program became so popular that ...

Naturally Evolving Architecture



- no time basis of data
- algorithmic differential
- levels of extraction
- external data
- no common source of data to begin with

[Inmon,2002]

Problems with this pattern of many extracting programs

- **Lack of Data credibility**

- **Productivity**

- **Inability to transform data into information**

The need for a different approach

Operational Systems (most - OLTP)

- **OLTP – On Line Transaction Processing**
  - **Systems that support the running activities of the organization**
  - **Examples:**
    - Point of sale in stores;
    - ATM and Bank operations
    - e-commerce (amazon, iTunes, etc)
- **Some characteristics:**
  - **Thousand of operations per second**
  - **Repeated operations dealing with small amounts of data (insert, update, remove)**
  - **Real Time**

DW and OLAP systems

- **OLAP – On Line Analytical Processing**
  - Systems that provide the users the necessary capabilities to analyze many and different aspects of organization activities and its performance.
  - Examples
    - How well certain product is selling in different regions? How well is the evolution in the market from its introduction?
    - Which are the top ten selling product in each region? and globally?
- **Some characteristics:**
  - Small number of queries (per day), when compared with OLTP systems
  - Large amount of data processed in each query, in order to obtain a small output.
  - It is hard to predict the queries and in general they are much more diverse, when compared with OLTP systems
  - Reading and processing data but no writing.

Analytic versus Operational

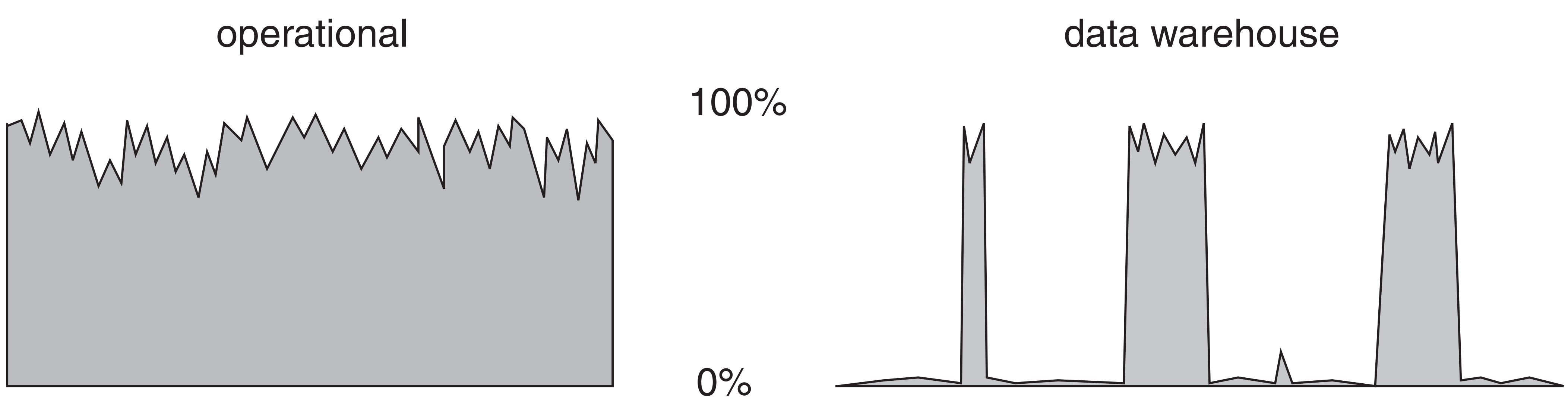
The users of an OLTP system are *running* the wheels of the organization.

The users of a data warehouse are *watching* the wheels of the organization

[Kimball,2002]

Analytic versus Operational - **Patterns of utilization**

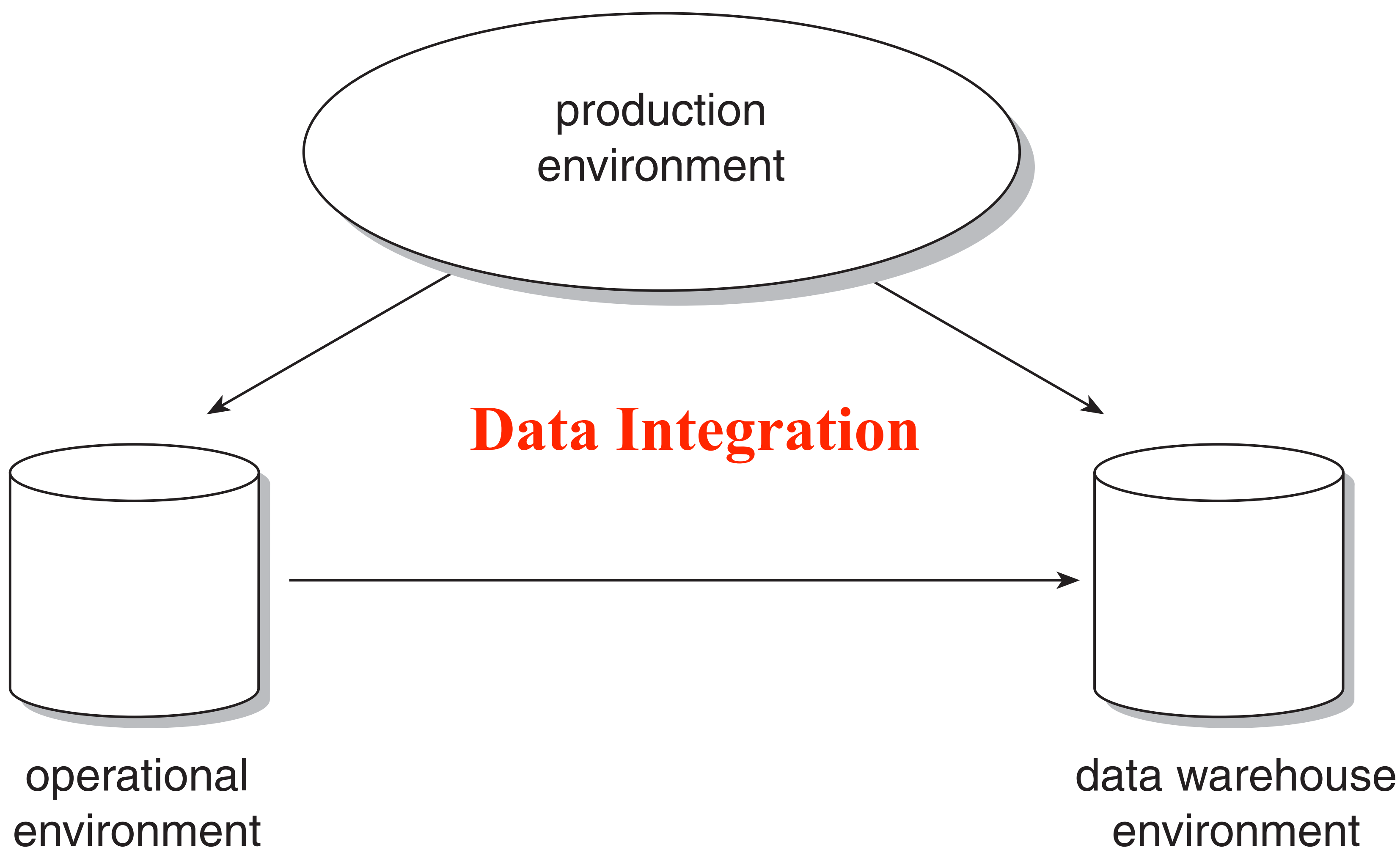
- **"The users of an OLTP system are *running* the wheels of the organization. The users of a data warehouse are *watching* the wheels of the organization" [Kimball]**



The different patterns of hardware utilization in the different environments.

Analytic versus Operational - **Separated Environments**

- **"The users of an OLTP system are *running* the wheels of the organization. The users of a data warehouse are *watching* the wheels of the organization" [Kimball]**



DW Reference Model

DW and OLAP systems

- **A data warehouse is an analytical database that is used as the foundation of a decision support system. It is designed for large volumes of read-only data, providing intuitive access to information that will be used in making decisions.**

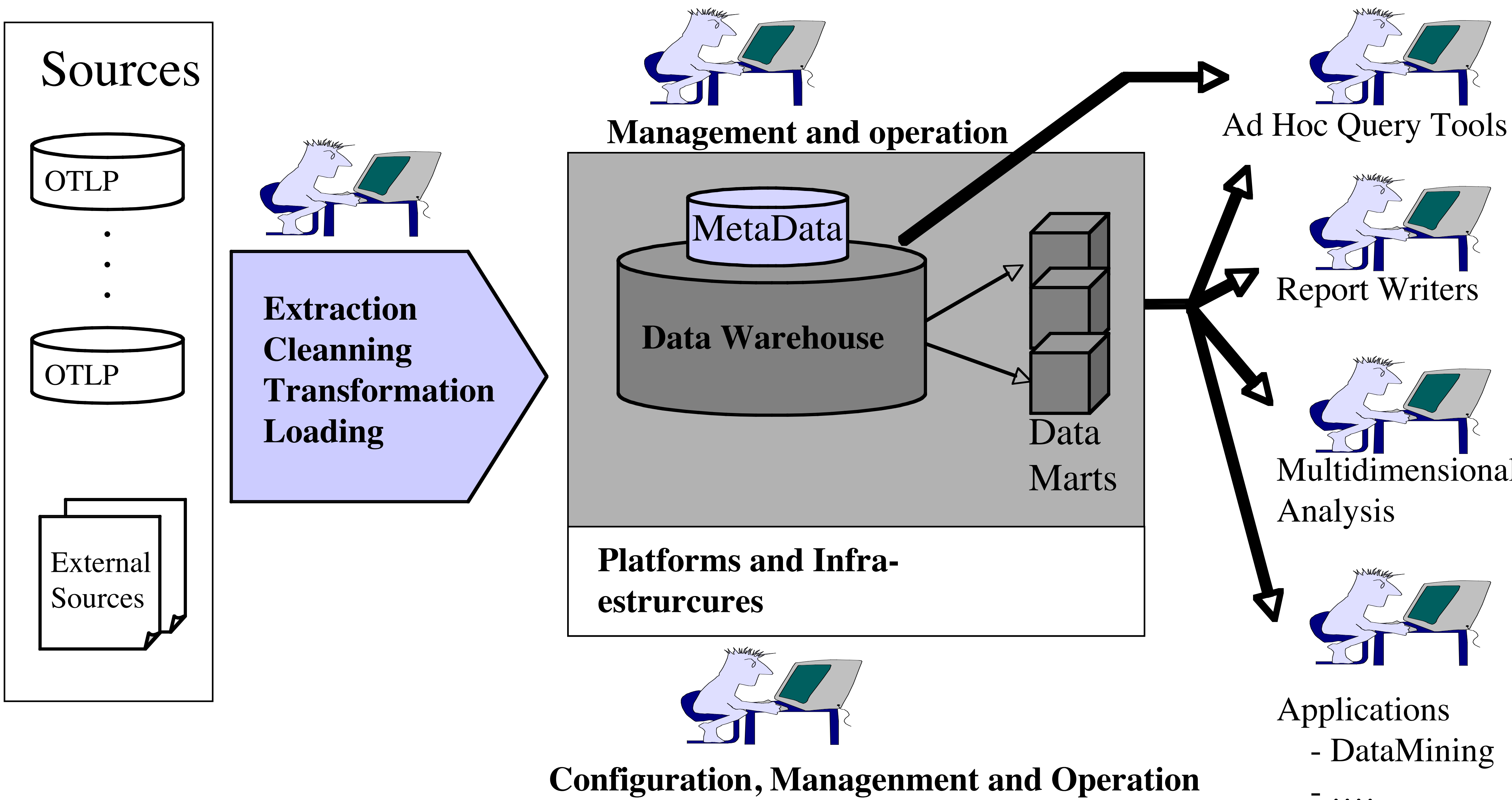
- **A data warehouse is created as ongoing commitment by the organization to ensure the appropriate data is available to the appropriate end user at the appropriate time**

[Vidette Poe, et all, 1997]

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.**

The Data Environment - Reference Model



Design Pattern for the DW. **Imnon School**

- **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 . **Imnon School**

- **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

+

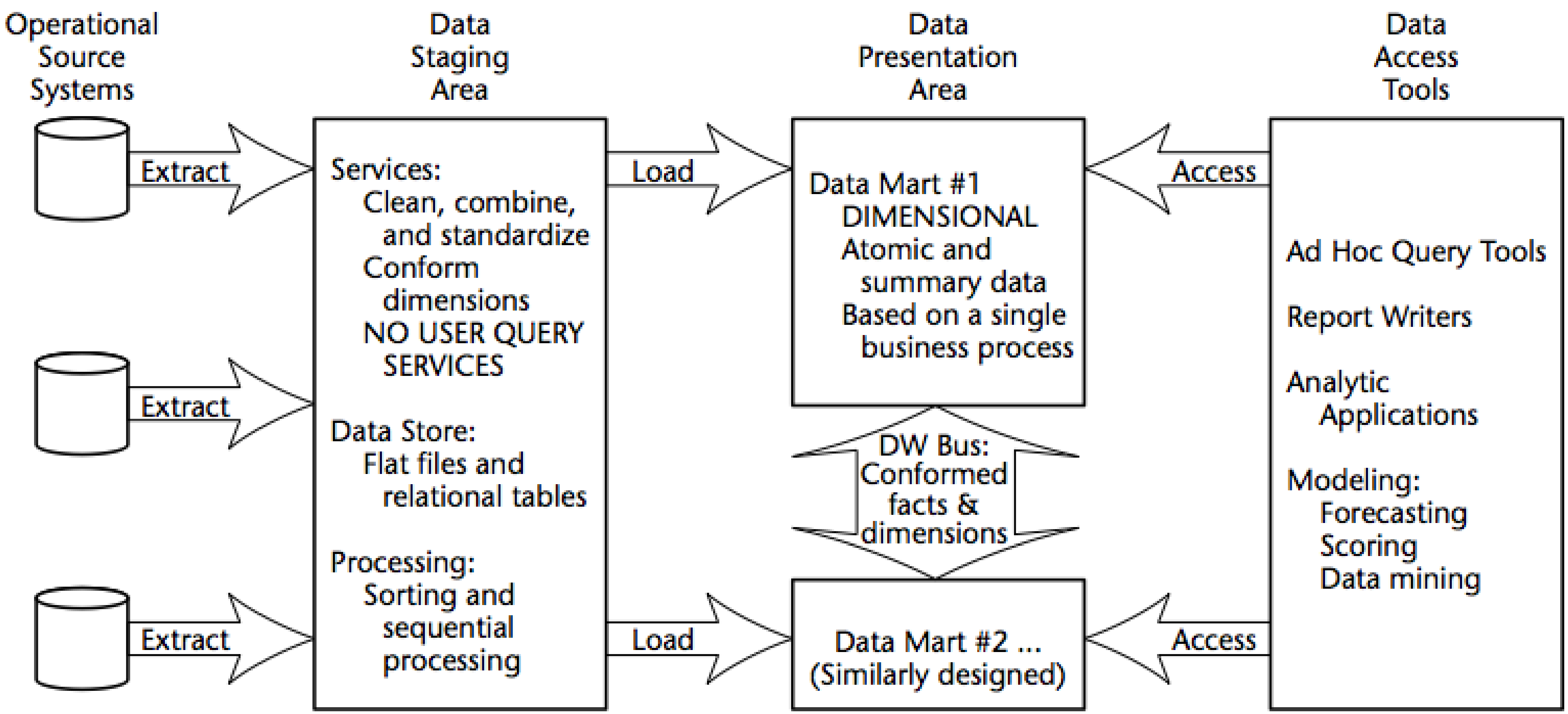
Historical Data

+

Structures Changes

Imnon School

Basic elements of the data warehouse (Kimball)



[Kimball, 2002]

Further Reading and Summary

Q&A

Further Reading and Summary

- **What you should know:**
  - The concept of Decision Support System, its evolution, the different types of DSS and the related Scientific areas.
  - DW and OLAP viewed as Data-Driven DSS. The justification to the actual importance of DW and OLAP in the DSS world.
  - Basic understand of the DW reference architecture
  - Fundamental differences from OLTP and OLAP systems, models, use, and users
  - Some analysis patterns used by OLAP users.

Introduction to Multidimensional Modeling

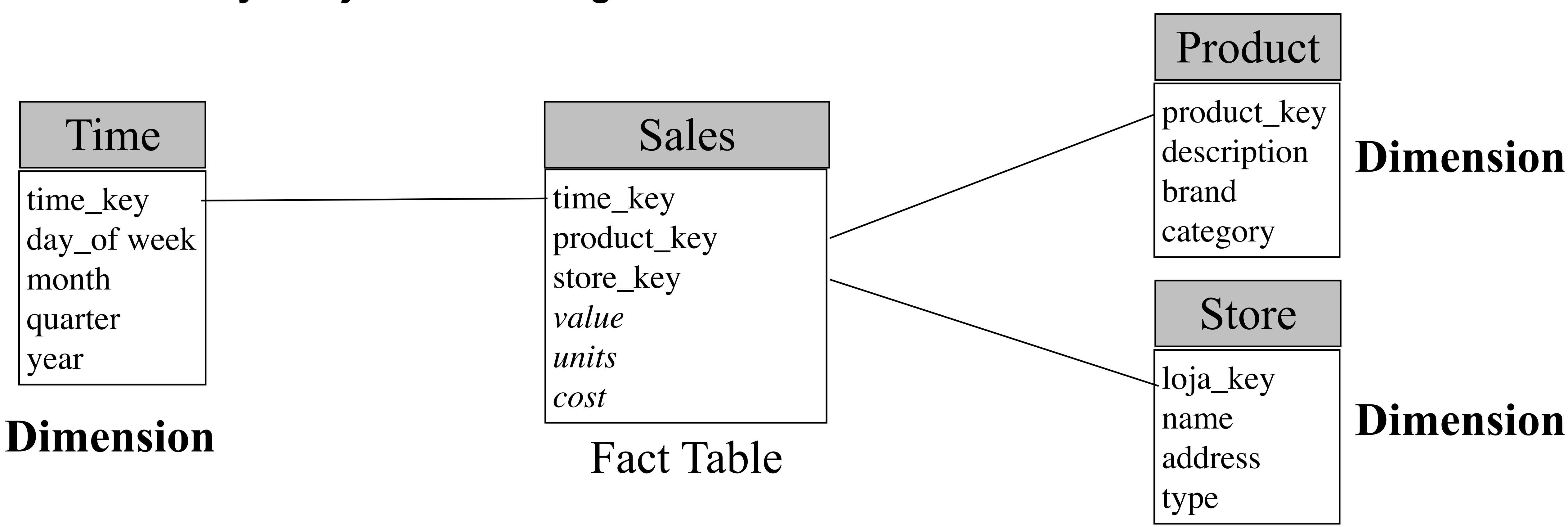
Multidimensional Approach

- **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



Fact Tables

- Numerical measures of process.
  - Continuous 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 of the query result**

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

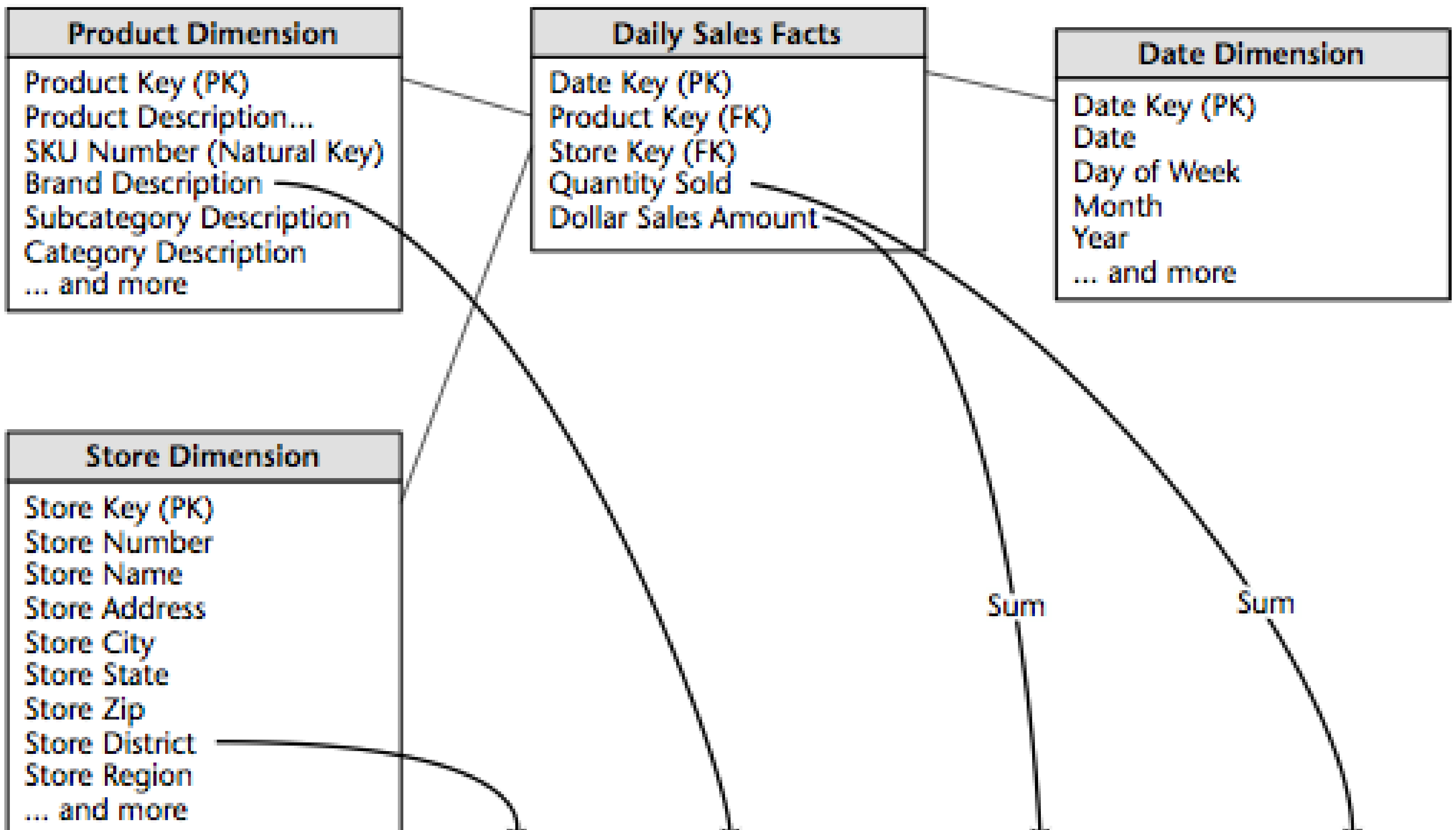
Dimension Tables

- Each dimension is defined by its **single primary key**, which serves as the basis for referential integrity with any given fact table to which it is joined.
  - The primary key defines the **dimension granularity**

- **Dimension attributes** serve as the primary source of **query constraints**, **groupings**, and **report labels**.
  - they are key to making the data warehouse usable and understandable. In many ways, the data warehouse is only as good as the dimension attributes
  - Dimension tables are the **entry points into the fact table**. Robust dimension attributes deliver **robust analytic slicing and dicing capabilities**. The dimensions implement the user interface to the data warehouse

[Kimball, 2002]

Dimension Tables - Attributes



District	Brand	Dollar Sales Amount	Quantity Sold
Atherton	Clean Fast	1,233	1,370
Atherton	More Power	2,239	2,035
Atherton	Zippy	848	707
Belmont	Clean Fast	2,097	2,330
Belmont	More Power	2,428	2,207
Belmont	Zippy	633	527

[Kimball, 2002]

Dimension Tables - attributes

- The best attributes are **textual** and **discrete**.
- **Sometimes** when we are designing a database it is **unclear** whether a numeric data field extracted from a production data source is a **fact** or **dimension attribute**.
  - If it takes on lots of values and participates in calculations - it is a fact
  - It is a discretely valued description that is more or less constant and participates in constraints - it is a dimensional attribute.
  - Occasionally, we can’t be certain of the classification. In such cases, it may be possible to model the data field either way, as a matter of designer’s prerogative.

[Kimball, 2002]

Dimension Tables - hierarchies

- Dimension tables often represent hierarchical relationships in the business.
  - In our sample product dimension table, products roll up into brands and then into categories.
- The hierarchical descriptive information is **stored redundantly**, in the spirit of **ease of use and query performance**. Dimension tables typically are highly de-normalized.
- The dimensions are usually quite small (less than 10 percent of the total data storage requirements). Since dimension tables typically are geometrically smaller than fact tables, improving storage efficiency by normalizing or snowflaking has virtually no impact on the overall database size.

[Kimball, 2002]

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

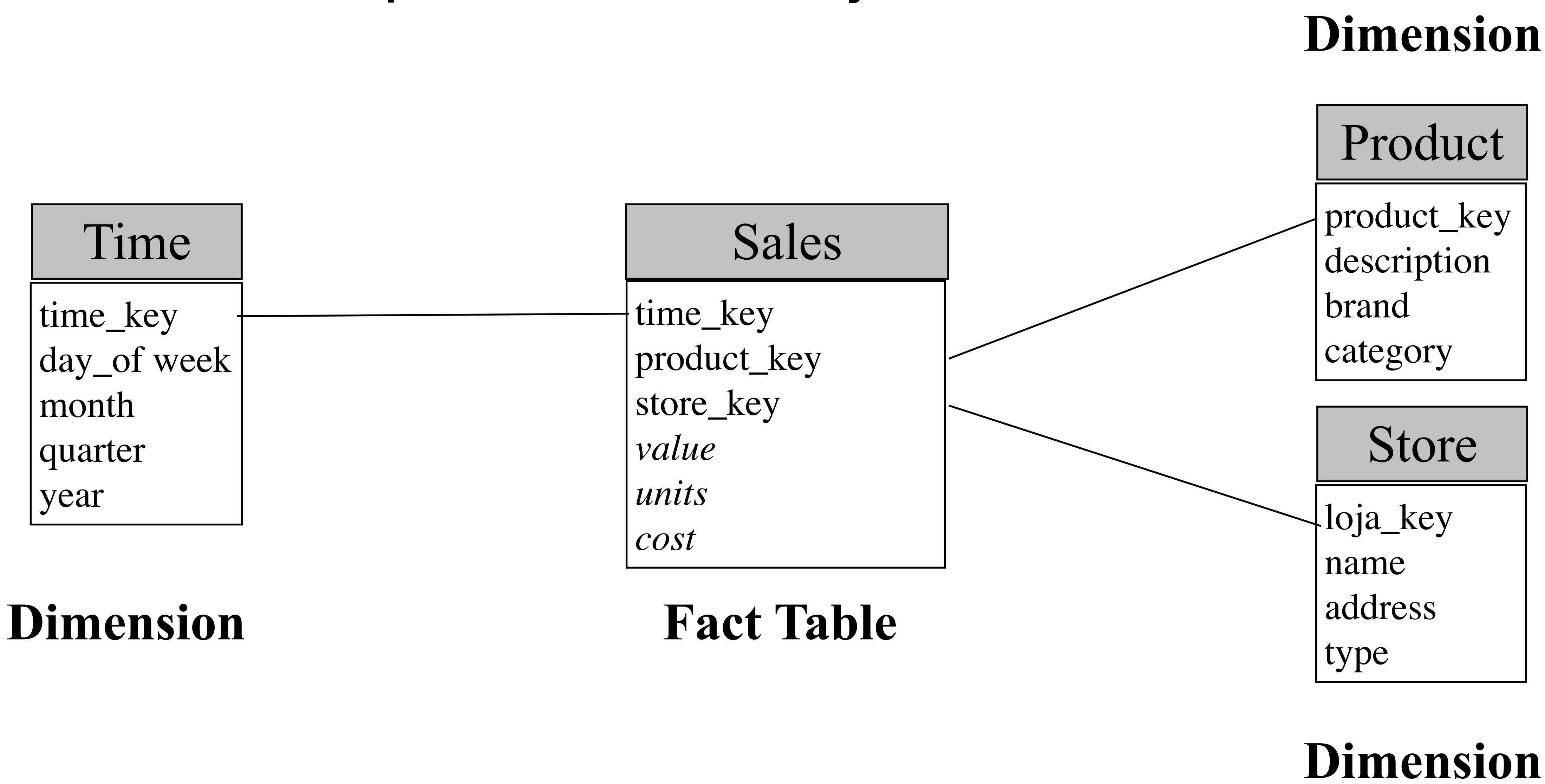
Metrics

Distinct values for the selected attribute

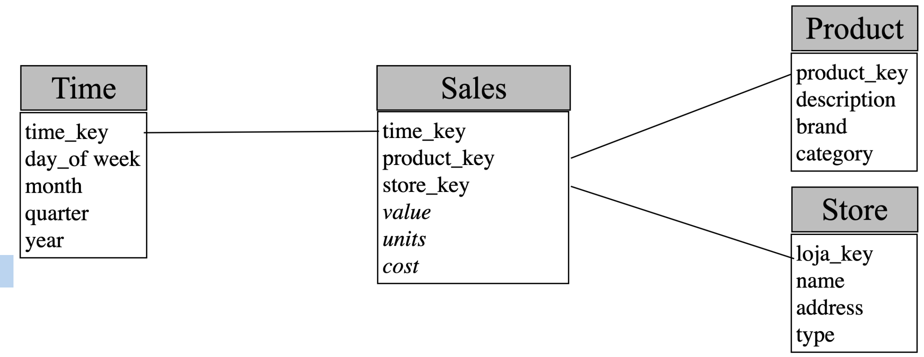
Textual Attribute of a Dimension

Querying a Star Schema

- **Data for the first quarter for all stores by brand**



Typical SQL query for OLAP



- **Data for the first quarter for all stores by brand**

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**

Dimensional attributes

Variables

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

Further Reading and Summary

Q&A

Further Reading and Summary

- **Readings**
  - (The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (Third Edition). pag 1 to 35.
- **What you should know:**
  - Understand the fundamental differences between OLTP and the analytical activities developed on the DW or on the Data Marts: data, access, users …
  - The basic idea of Kimball school: the DW is a collection of Multidimensional Data Marts that are built incrementally and are made compatible

Further Reading and Summary

- **What you should know:**
  - The basic OLAP vocabulary: dimensions, mesures, aggregation, slice, drill-down and drill-up
  - The basic building blocks for multidimensional models: Dimensions and Facts.
  - The basic components of a Multidimensional query
  - The meaning of “browsing” dimensions
  - The presentation results and its relation to the OLAP operations

Further Reading and Summary

- **What you should know:**
  - The basic components of a Multidimensional query
  - The meaning of “browsing” dimensions
  - The presentation results and its relation to the OLAP operations