Data Warehousing: Data Models and OLAP operations

By
Kishore Jaladi
kishorejaladi@yahoo.com

Topics Covered

1. Understanding the term "Data Warehousing"
2. Three-tier Decision Support Systems
3. Approaches to OLAP servers
4. Multi-dimensional data model
5. ROLAP
6. MOLAP
7. HOLAP
8. Which to choose: Compare and Contrast
9. Conclusion

Understanding the term Data Warehousing

- **Data Warehouse:**

  The term Data Warehouse was coined by Bill Inmon in 1990, which he defined in the following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". He defined the terms in the sentence as follows:

- **Subject Oriented:**

  Data that gives information about a particular subject instead of about a company's ongoing operations.

- **Integrated:**

  Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

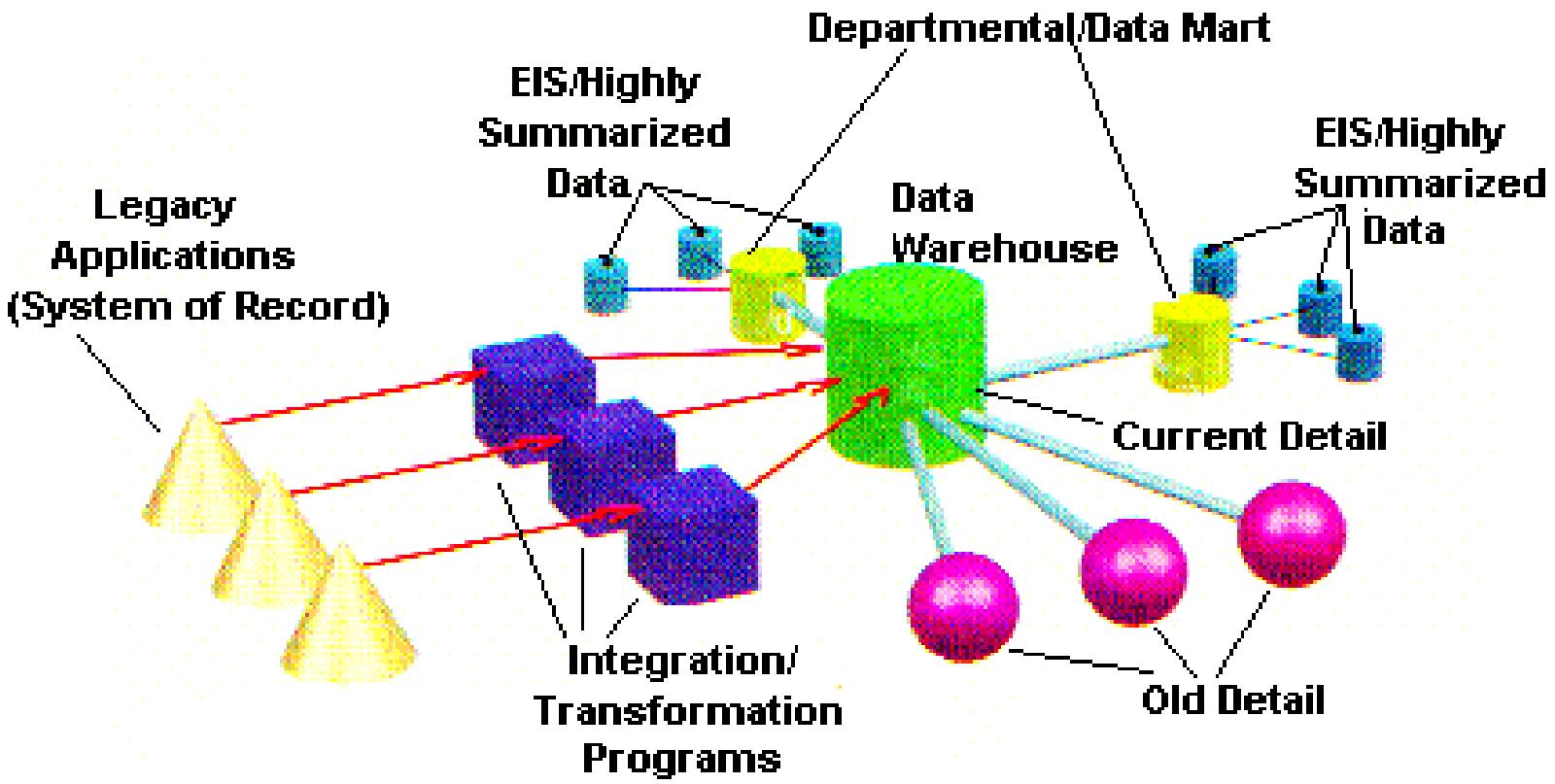
- **Time-variant:**

  All data in the data warehouse is identified with a particular time period.

- **Non-volatile**

  Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

Data Warehouse Architecture



Other important terminology

- **Enterprise Data warehouse**

  collects all information about subjects (*customers,products,sales,assets, personnel*) that span the entire organization

- **Data Mart**

  Departmental subsets that focus on selected subjects

- **Decision Support System (DSS)**

  Information technology to help the knowledge worker (executive, manager, analyst) make faster & better decisions

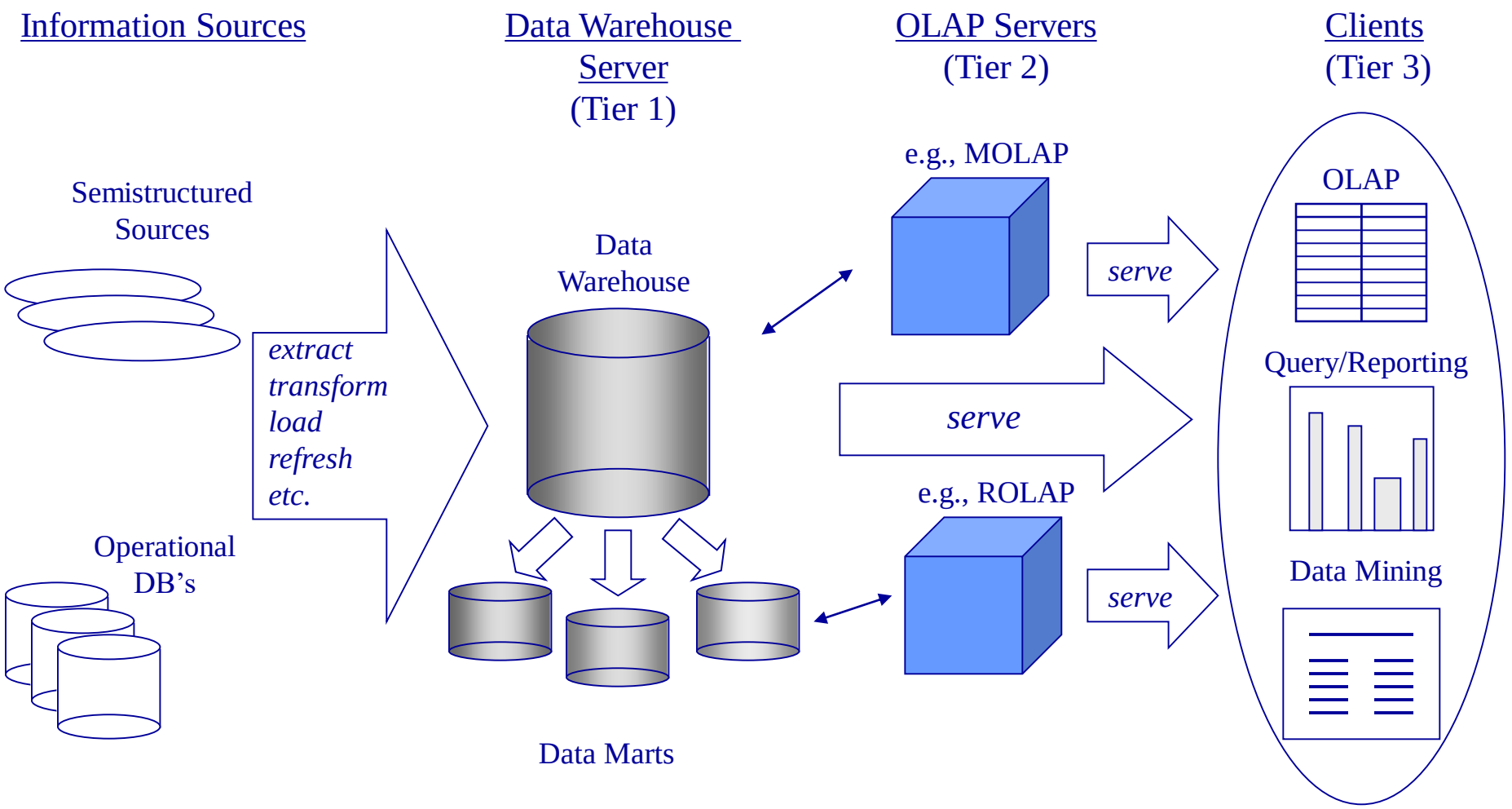
- **Online Analytical Processing (OLAP)**

  an element of decision support systems (DSS)

Three-Tier Decision Support Systems

- Warehouse database server
  - Almost always a relational DBMS, rarely flat files
- OLAP servers
  - Relational OLAP (ROLAP): extended relational DBMS that maps operations on multidimensional data to standard relational operators
  - Multidimensional OLAP (MOLAP): special-purpose server that directly implements multidimensional data and operations
- Clients
  - Query and reporting tools
  - Analysis tools
  - Data mining tools

The Complete Decision Support System



Approaches to OLAP Servers

Three possibilities for OLAP servers

(1) Relational OLAP (ROLAP)

- Relational and specialized relational DBMS to store and manage warehouse data
- OLAP middleware to support missing pieces

(2) Multidimensional OLAP (MOLAP)

- Array-based storage structures
- Direct access to array data structures

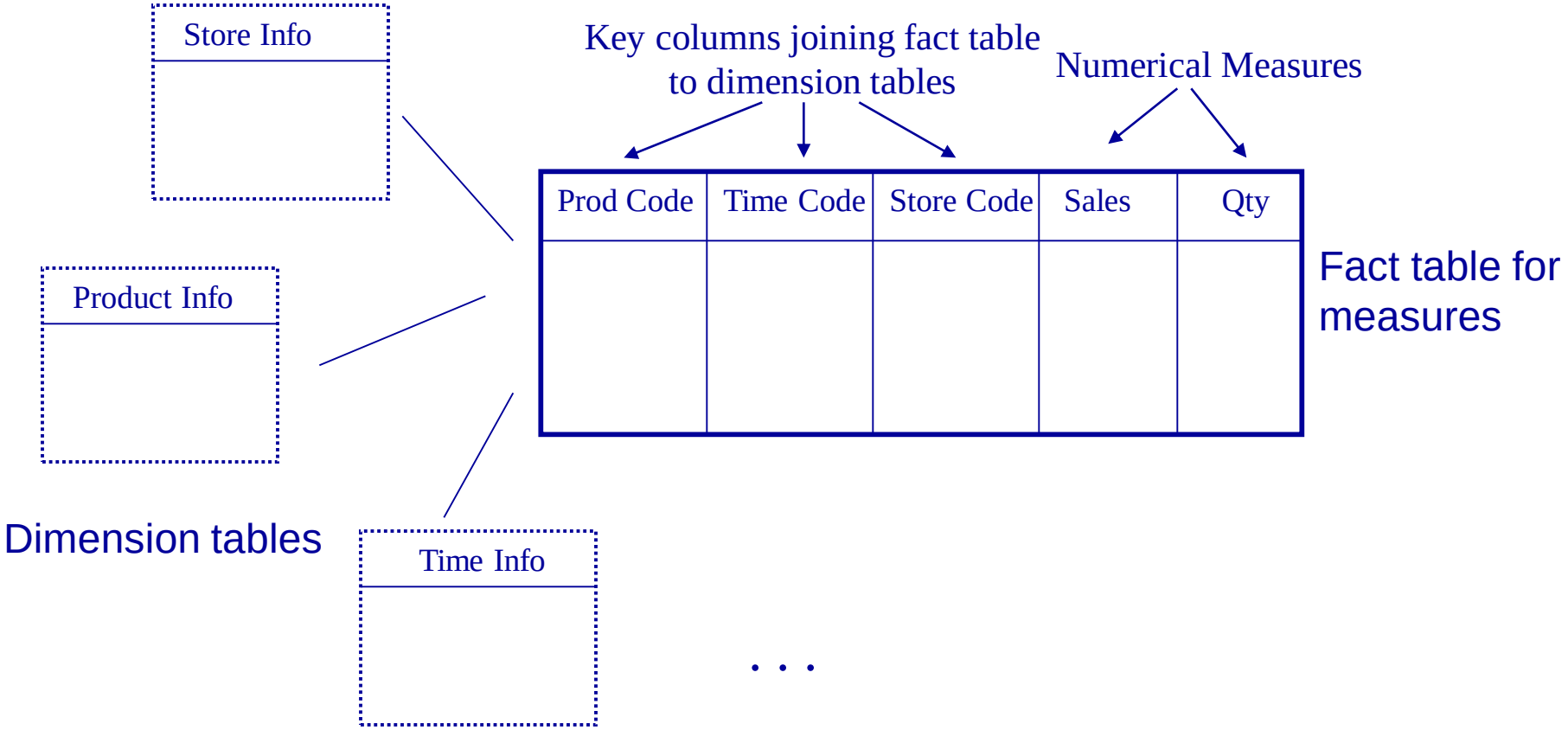
(3) Hybrid OLAP (HOLAP)

- Storing detailed data in RDBMS
- Storing aggregated data in MDBMS
- User access via MOLAP tools

The Multi-Dimensional Data Model

"Sales by product line over the past six months"

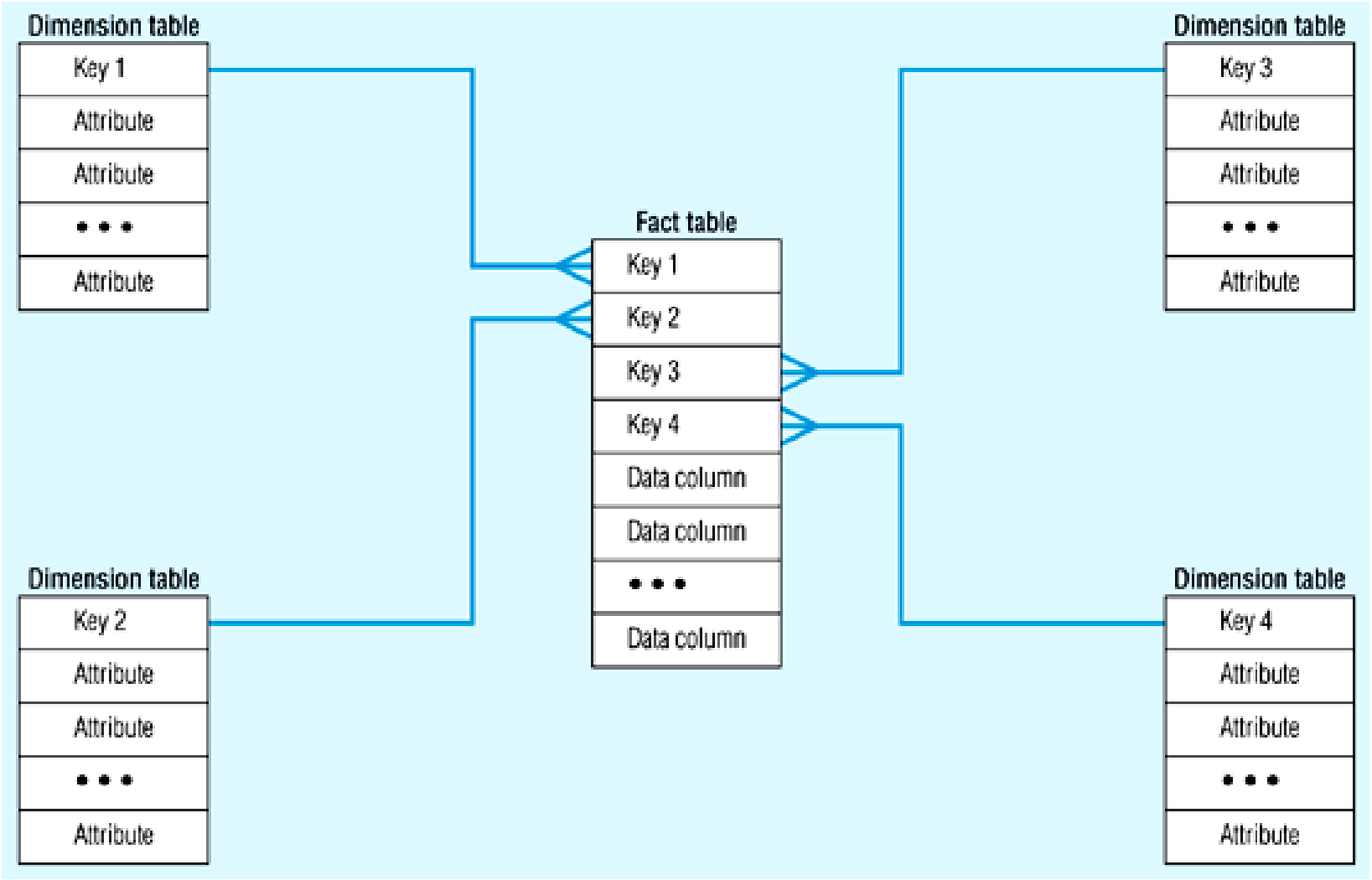
"Sales by store between 1990 and 1995"



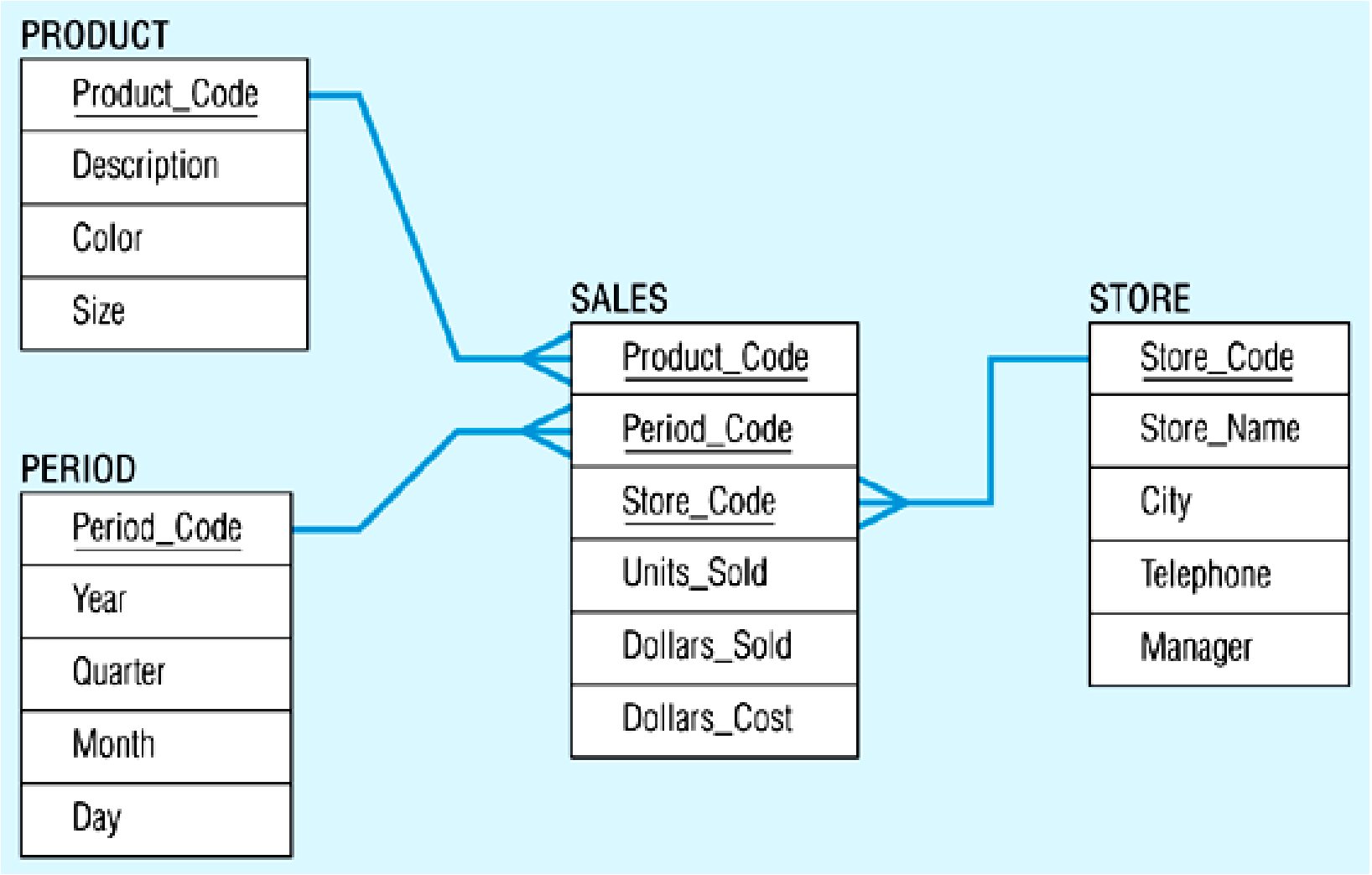
ROLAP: Dimensional Modeling Using Relational DBMS

- Special schema design: *star, snowflake*
- Special indexes: bitmap, multi-table join
- Proven technology (relational model, DBMS), tend to outperform specialized MDDB especially on large data sets
- Products
  - IBM DB2, Oracle, Sybase IQ, RedBrick, Informix

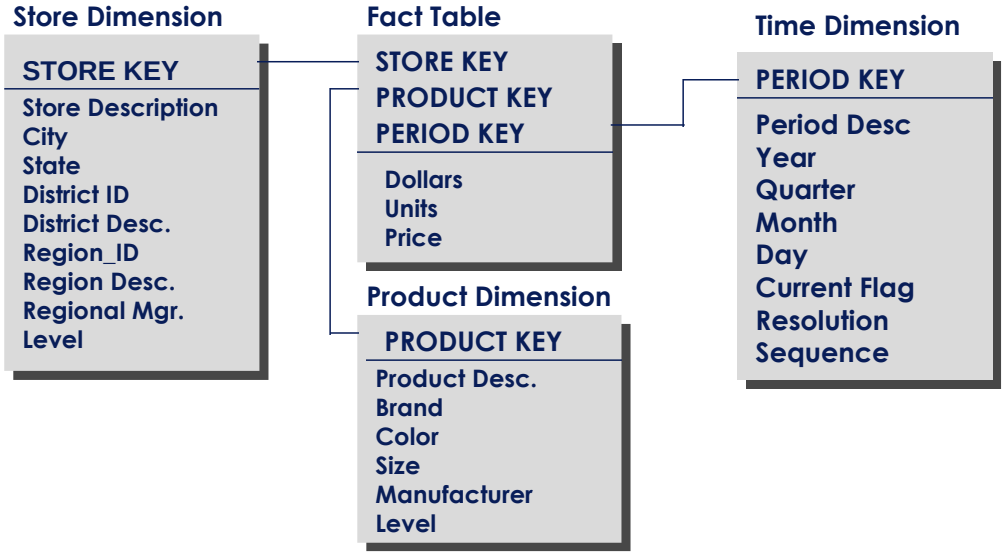
Star Schema (in RDBMS)



Star Schema Example



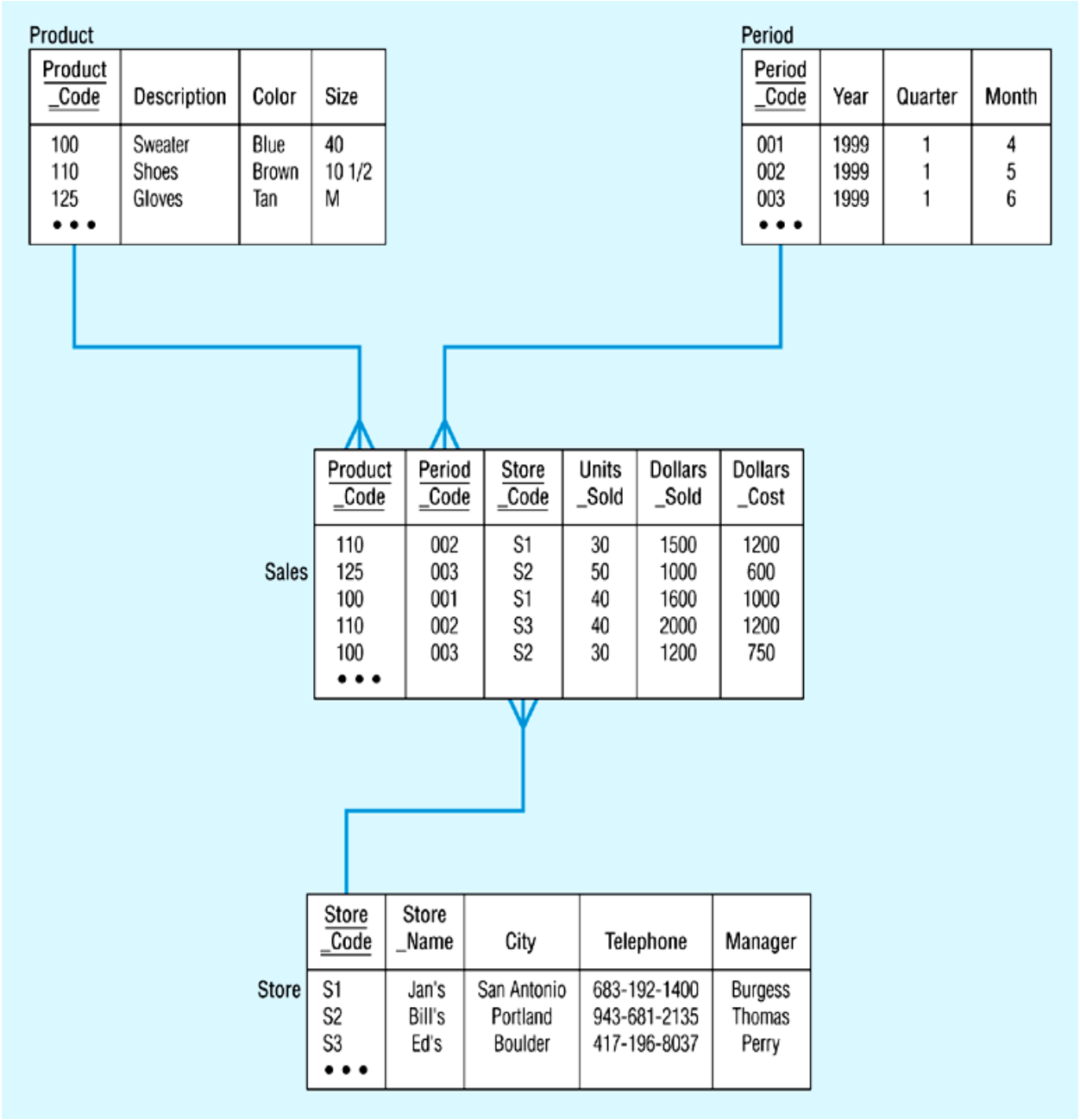
The "Classic" Star Schema



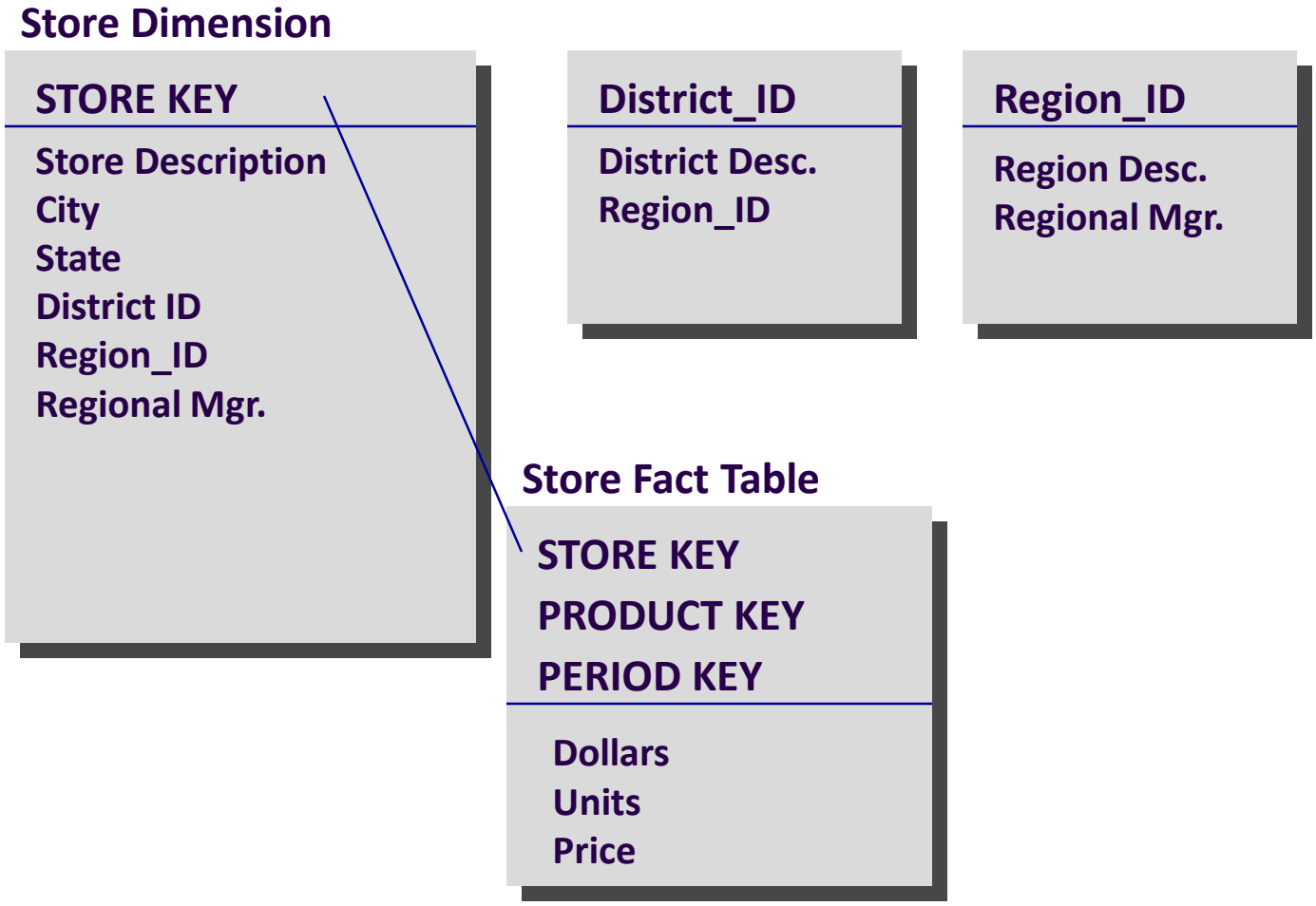
- A single fact table, with detail and summary data
- Fact table primary key has only one key column per dimension
- Each key is generated
- Each dimension is a single table, highly de-normalized

Benefits: Easy to understand, easy to define hierarchies, reduces # of physical joins, low maintenance, very simple metadata

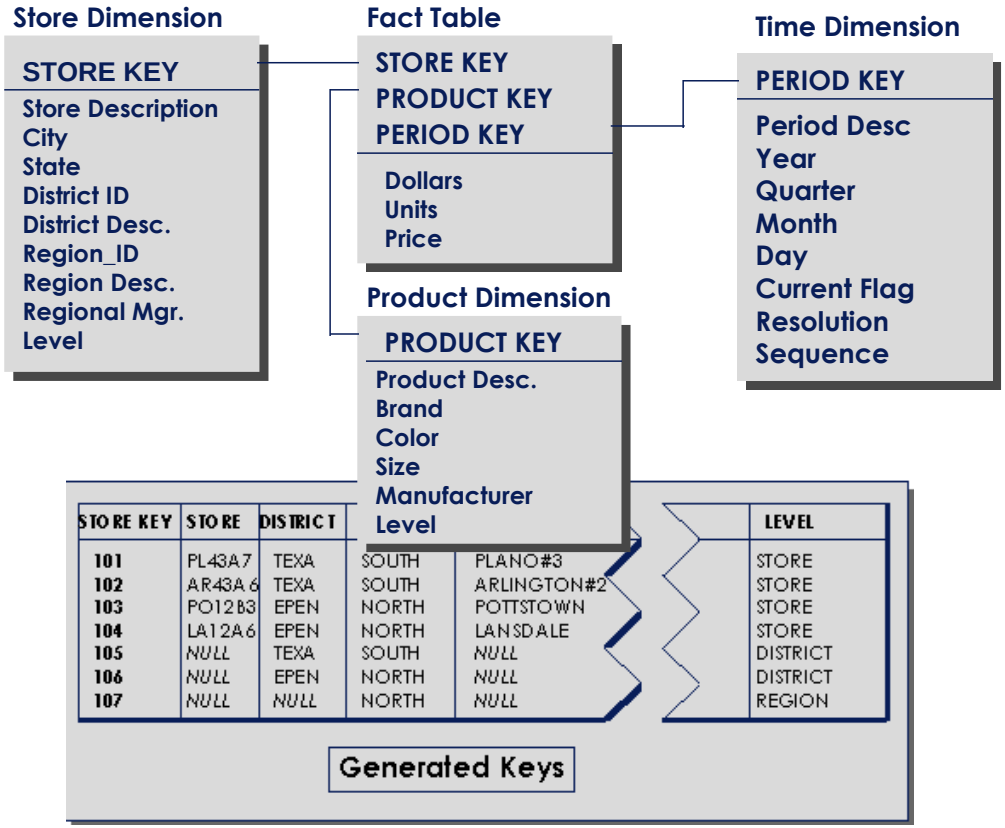
Star Schema with Sample Data



The "Snowflake" Schema

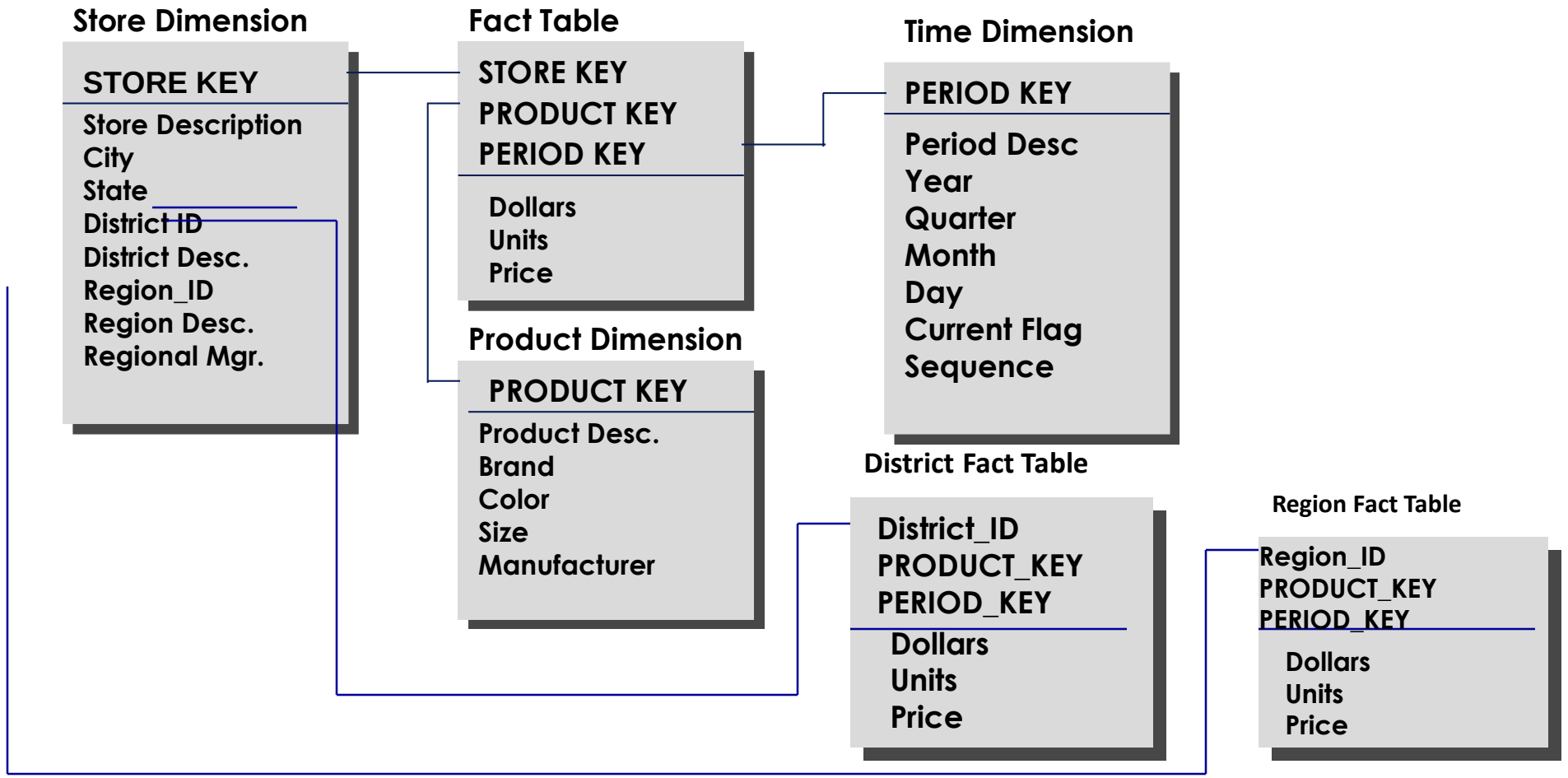


Aggregation in a Single Fact Table

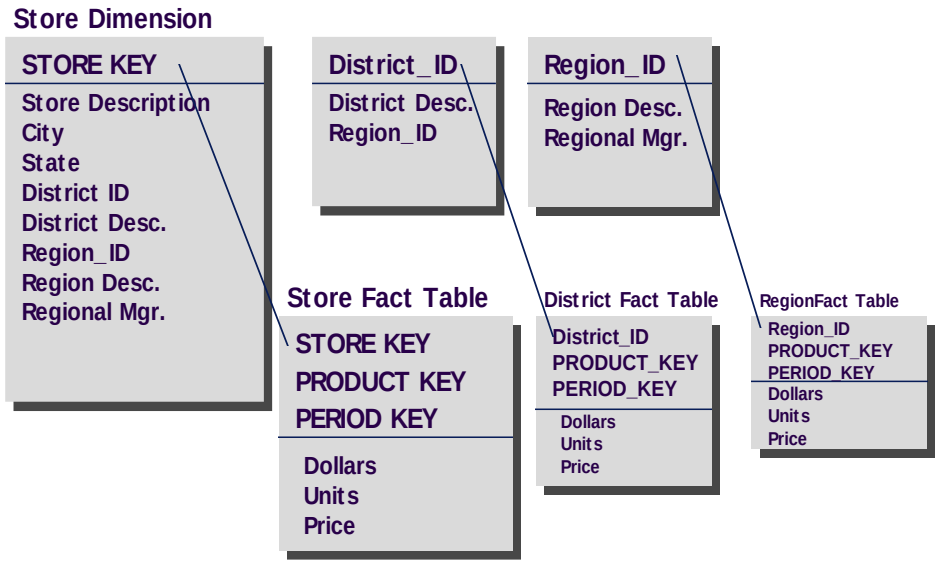


Drawbacks: Summary data in the fact table yields poorer performance for summary levels, huge dimension tables a problem

The "Fact Constellation" Schema



Aggregations using "Snowflake" Schema and Multiple Fact Tables

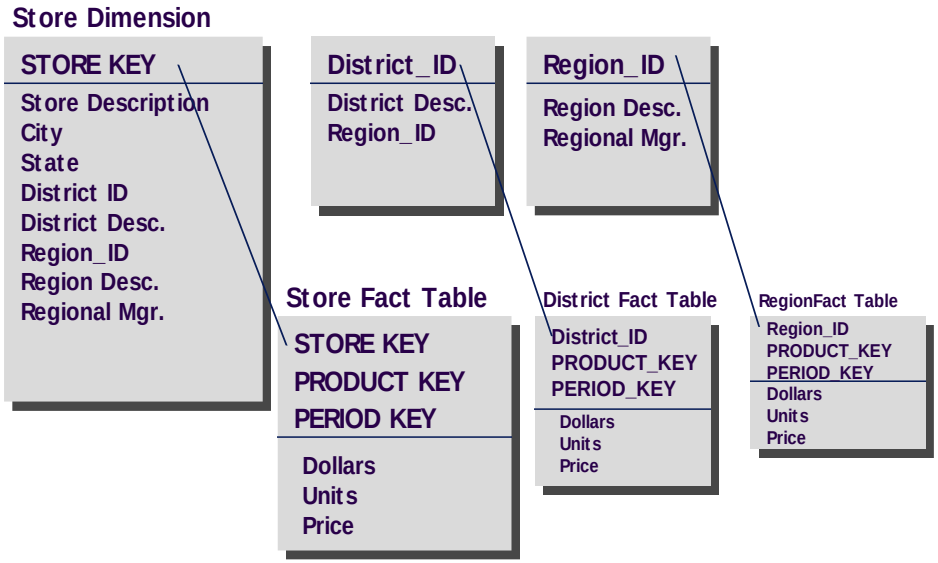


- No **LEVEL** in dimension tables
- Dimension tables are normalized by decomposing at the attribute level
- Each dimension table has one key for each level of the dimensionís hierarchy
- The lowest level key joins the dimension table to both the fact table and the lower level attribute table

How does it work?

The best way is for the query to be built by understanding which summary levels exist, and finding the proper snowflaked attribute tables, constraining there for keys, then selecting from the fact table.

Aggregation Contd …



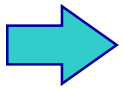
Advantage: Best performance when queries involve aggregation

Disadvantage: Complicated maintenance and metadata, explosion in the number of tables in the database

Aggregates

- Add up amounts for day 1
- In SQL: SELECT sum(amt) FROM SALE
  WHERE date = 1

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

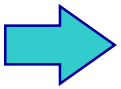


81

Aggregates

- Add up amounts by day
- In SQL: SELECT date, sum(amt) FROM SALE GROUP BY date

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

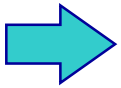


ans	date	sum
	1	81
	2	48

Another Example

- Add up amounts by day, product
- In SQL: SELECT date, sum(amt) FROM SALE GROUP BY date, prodId

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4



sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48

rollup →

← drill-down

Points to be noticed about ROLAP

- Defines complex, multi-dimensional data with simple model
- Reduces the number of joins a query has to process
- Allows the data warehouse to evolve with rel. low maintenance
- Can contain both detailed and summarized data.
- ROLAP is based on familiar, proven, and already selected technologies.

BUT!!!

- SQL for multi-dimensional manipulation of calculations.

MOLAP: Dimensional Modeling Using the Multi Dimensional Model

- MDDB: a special-purpose data model
- Facts stored in multi-dimensional arrays
- Dimensions used to index array
- Sometimes on top of relational DB
- Products
  - Pilot, Arbor Essbase, Gentia

The MOLAP Cube

Fact table view:

sale	prodId	storeId	amt
	p1	s1	12
	p2	s1	11
	p1	s3	50
	p2	s2	8

Multi-dimensional cube:

	s1	s2	s3
p1	12		50
p2	11	8	

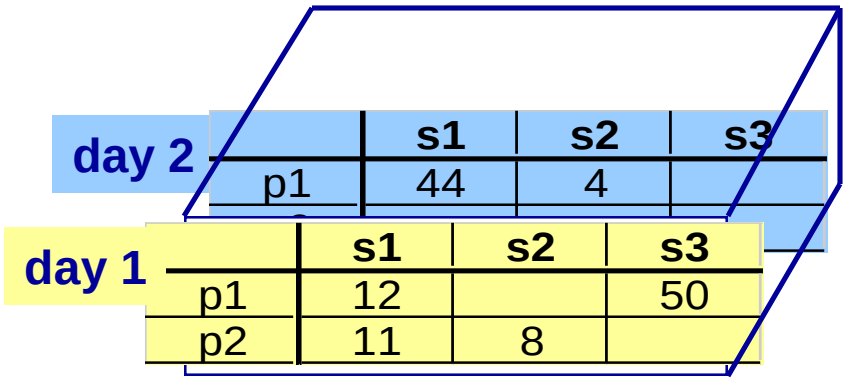
dimensions = 2

3-D Cube

Fact table view:

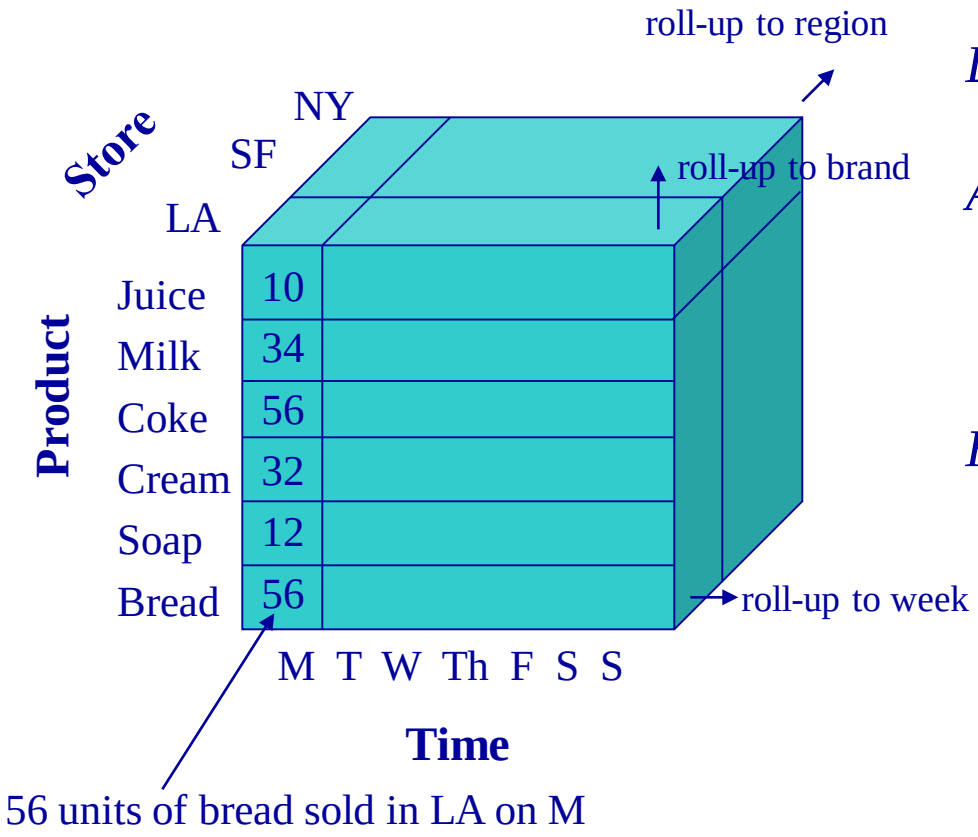
sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

Multi-dimensional cube:



dimensions = 3

Example



Dimensions:

Time, Product, Store

Attributes:

Product (upc, price, …)

Store …

…

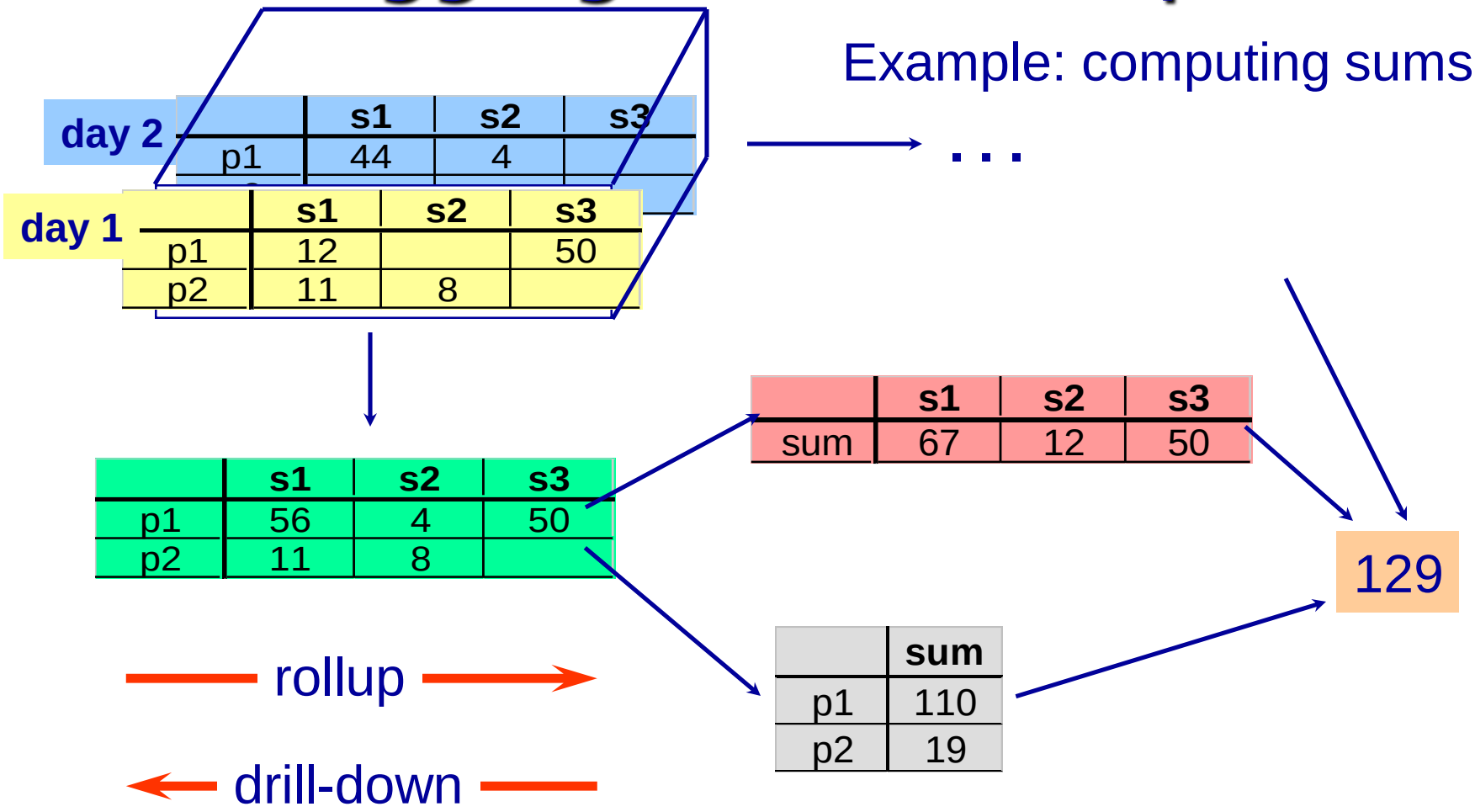
Hierarchies:

Product → Brand → …

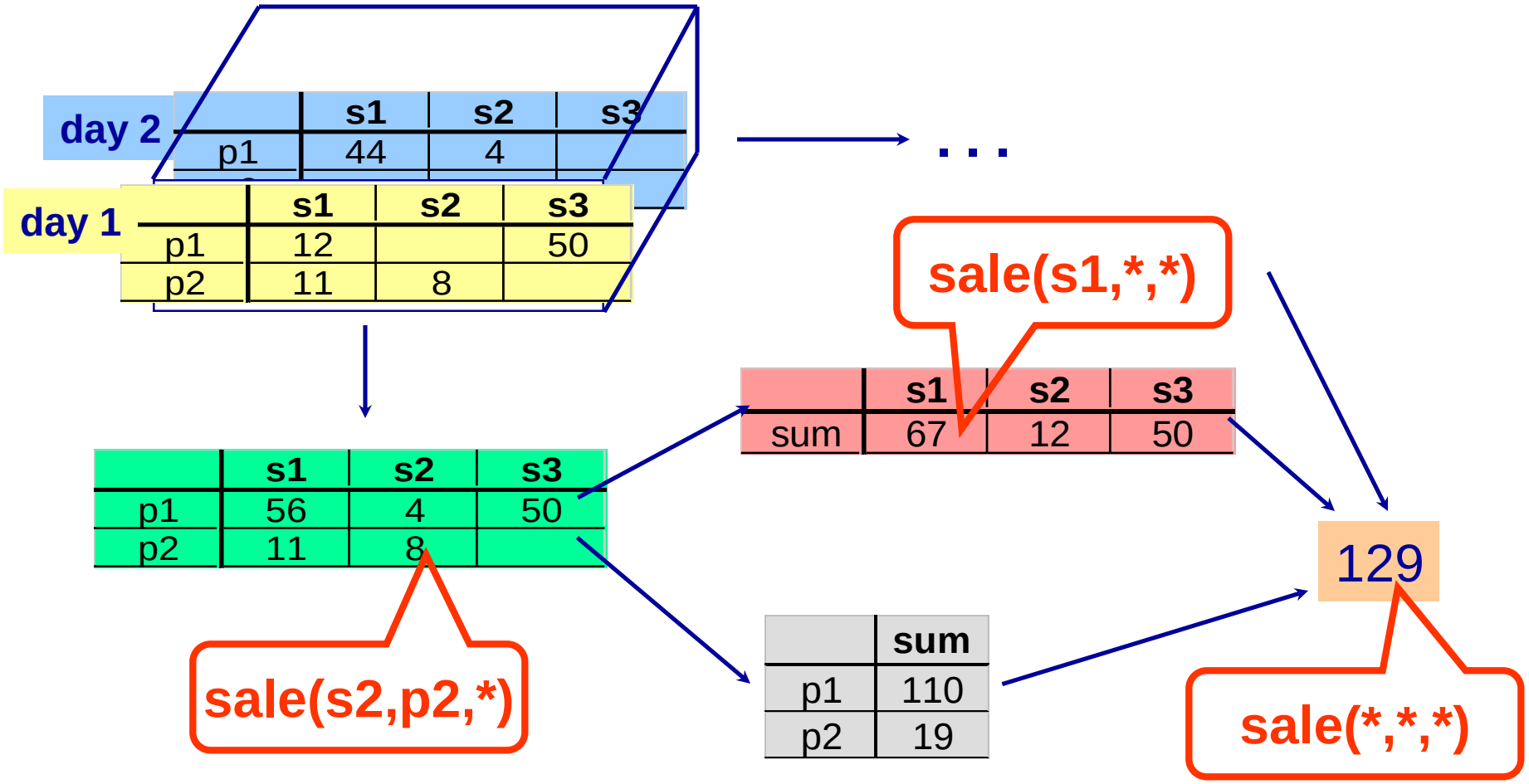
Day → Week → Quarter

Store → Region → Country

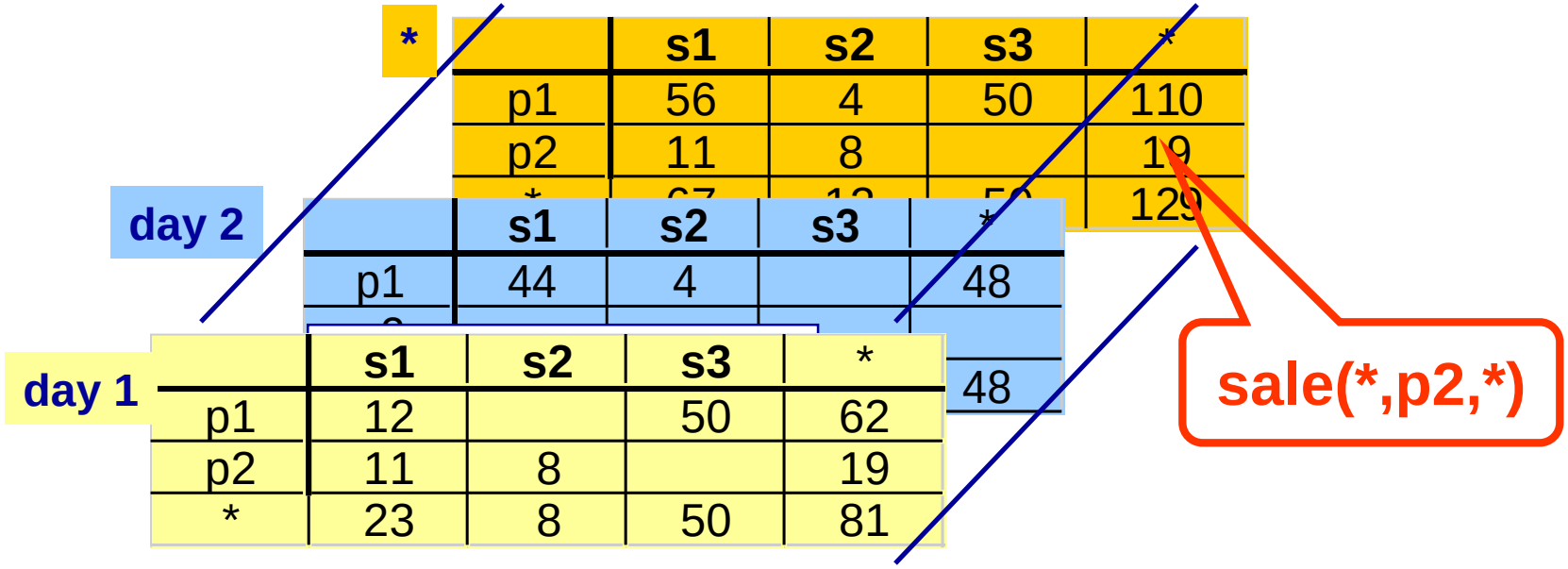
Cube Aggregation: Roll-up



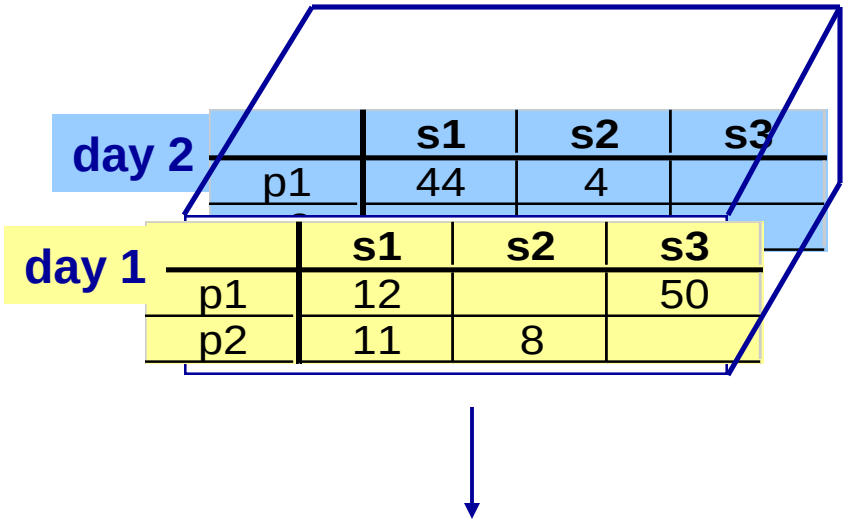
Cube Operators for Roll-up



Extended Cube



Aggregation Using Hierarchies



	region A	region B
p1	56	54
p2	11	8

store

region

country

(store s1 in Region A;
stores s2, s3 in Region B)

Points to be noticed about MOLAP

- Pre-calculating or pre-consolidating transactional data improves speed.

BUT

Fully pre-consolidating incoming data, MDDs require an enormous amount of overhead both in processing time and in storage. An input file of 200MB can easily expand to 5GB

MDDs are great candidates for the <50GB department data marts.

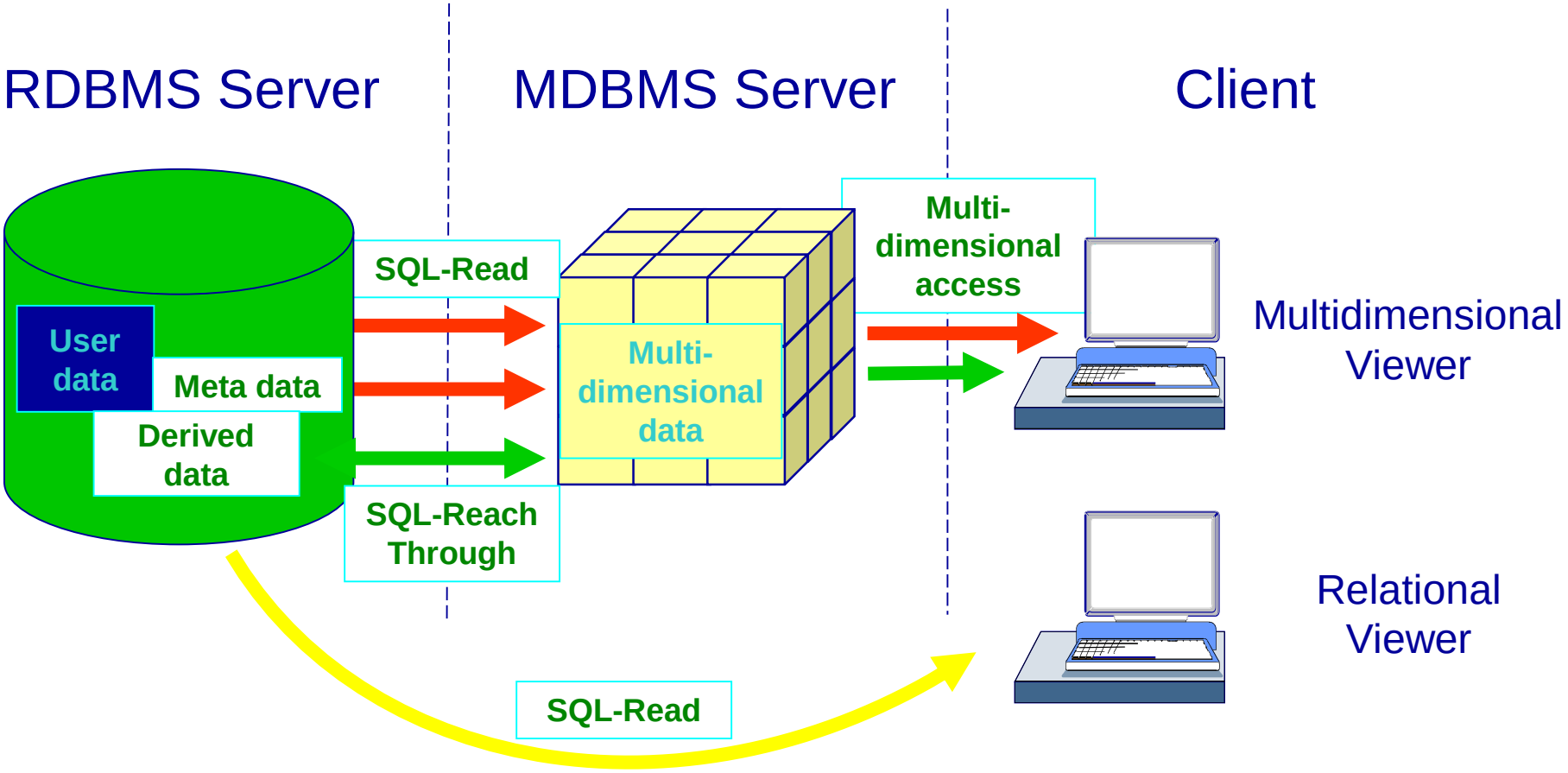
- Rolling up and Drilling down through aggregate data.

- With MDDs, application design is essentially the definition of dimensions and calculation rules, while the RDBMS requires that the database schema be a star or snowflake.

Hybrid OLAP (HOLAP)

- HOLAP = Hybrid OLAP:
  - Best of both worlds
  - Storing detailed data in RDBMS
  - Storing aggregated data in MDBMS
  - User access via MOLAP tools

Data Flow in HOLAP



When deciding which technology to go for, consider:

1) Performance:

- How fast will the system appear to the end-user?
- MDD server vendors believe this is a key point in their favor.

2) Data volume and scalability:

- While MDD servers can handle up to 50GB of storage, RDBMS servers can handle hundreds of gigabytes and terabytes.

An experiment with Relational and the Multidimensional models on a data set

The analysis of the author's example illustrates the following differences between the best Relational alternative and the Multidimensional approach.

	relational	Multi-dimensional	Improvement
Disk space requirement (Gigabytes)	17	10	1.7
Retrieve the corporate measures Actual Vs Budget, by month (I/O's)	240	1	240
Calculation of Variance Budget/Actual for the whole database (I/O time in hours)	237	2*	110*

* This may include the calculation of many other derived data without any additional I/O.

Reference: http://dimlab.usc.edu/csci599/Fall2002/paper/I2_P064.pdf

What-if analysis

IF

A. You require write access
B. Your data is under 50 GB
C. Your timetable to implement is 60-90 days
D. Lowest level already aggregated
E. Data access on aggregated level
F. You're developing a general-purpose application for inventory movement or assets management

THEN

Consider an **MDD /MOLAP** solution for your data mart

IF

A. Your data is over 100 GB
B. You have a "read-only" requirement
C. Historical data at the lowest level of granularity
D. Detailed access, long-running queries
E. Data assigned to lowest level elements

THEN

Consider an **RDBMS/ROLAP** solution for your data mart.

IF

A. OLAP on aggregated and detailed data
B. Different user groups
C. Ease of use and detailed data

THEN

Consider an **HOLAP** for your data mart

Examples

- ROLAP
  - Telecommunication startup: call data records (CDRs)
  - ECommerce Site
  - Credit Card Company
- MOLAP
  - Analysis and budgeting in a financial department
  - Sales analysis
- HOLAP
  - Sales department of a multi-national company
  - Banks and Financial Service Providers

Tools available

- ROLAP:
  - ORACLE 8i
  - ORACLE Reports; ORACLE Discoverer
  - ORACLE Warehouse Builder
  - Arbors Software's Essbase
- MOLAP:
  - ORACLE Express Server
  - ORACLE Express Clients (C/S and Web)
  - MicroStrategy's DSS server
  - Platinum Technologies' Plantinum InfoBeacon
- HOLAP:
  - ORACLE 8i
  - ORACLE Express Serve
  - ORACLE Relational Access Manager
  - ORACLE Express Clients (C/S and Web)

Conclusion

- ROLAP: RDBMS -> star/snowflake schema
- MOLAP: MDD -> Cube structures
- ROLAP or MOLAP: Data models used play major role in performance differences
- MOLAP: for summarized and relatively lesser volumes of data (10-50GB)
- ROLAP: for detailed and larger volumes of data
- Both storage methods have strengths and weaknesses
- The choice is requirement specific, though currently data warehouses are predominantly built using RDBMSs/ROLAP.

References

- **OLAP, Relational, and Multidimensional Database Systems,** by George Colliat, Arbor Software Corporation
- **Data warehousing Services**, *Data Mining & Analysis, LLC*
- http://www.cs.man.ac.uk/~franconi/teaching/2001/CS636/CS636-olap.ppt
  - **Data Warehouse Models and OLAP Operations**, by Enrico Franconi
- ROLAP, MOLAP, HOLAP: How to determine which to technology is appropriate, by Holger Frietch, PROMATIS Corporation