Database management II.

Data warehousing

Gergely Lukács
Pázmány Péter Catholic University
Faculty of Information Technology
Budapest, Hungary
lukacs@itk.ppke.hu

INTRODUCTION

Motivation

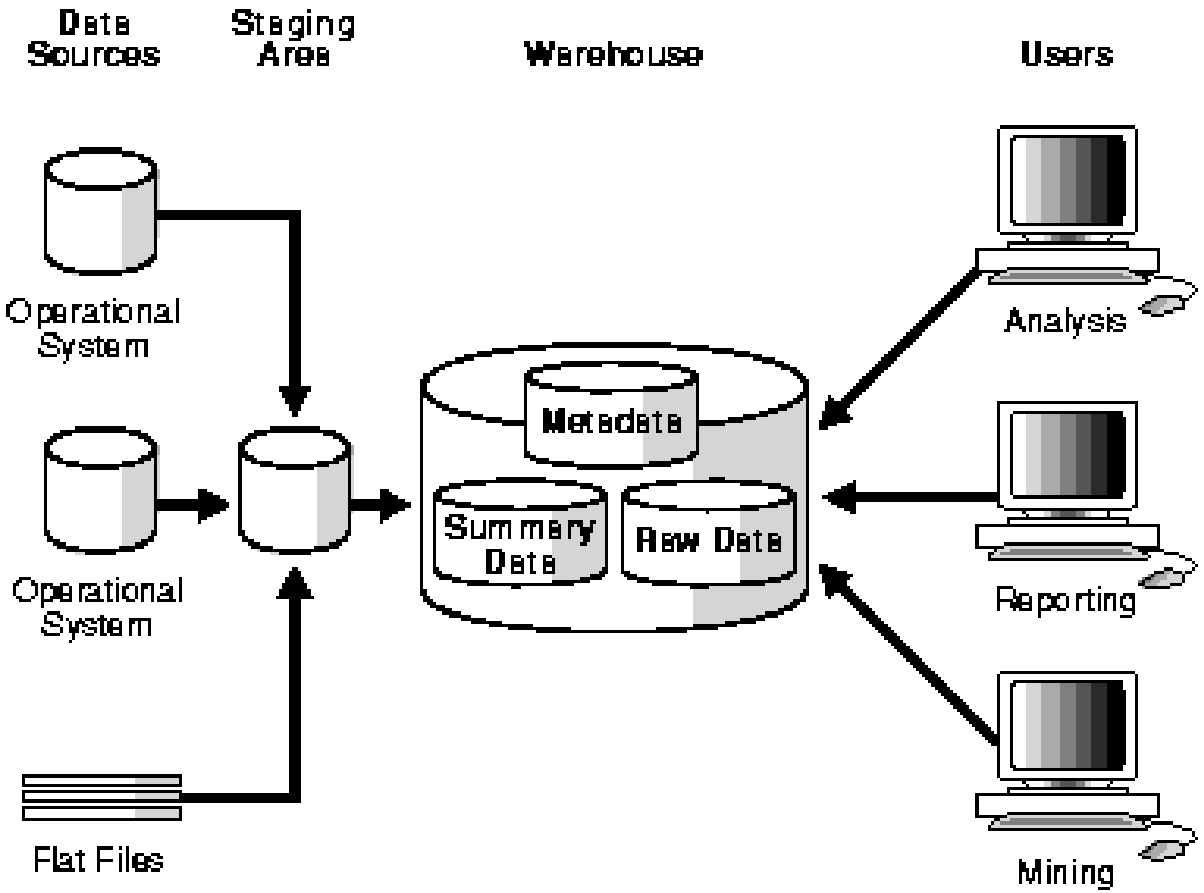
- Up to now:
  - Supporting individual, standard operations (buying, transferring money, lending a book in a libary,…)
  - Standardized business processes
  - Small transactions with data modifications
  - folyamatok
- Now (incl. Tableau)
  - Supporting strategic decisions
  - When should the library be open?
  - Which books shell we buy?
  - Trends
    - How does this change in time, space and over organisational units

Definition

- **A data warehouse is a copy of transaction data specifically structured for query and analysis. (Ralph Kimball)**

Definition - 2 (Gartner Group)

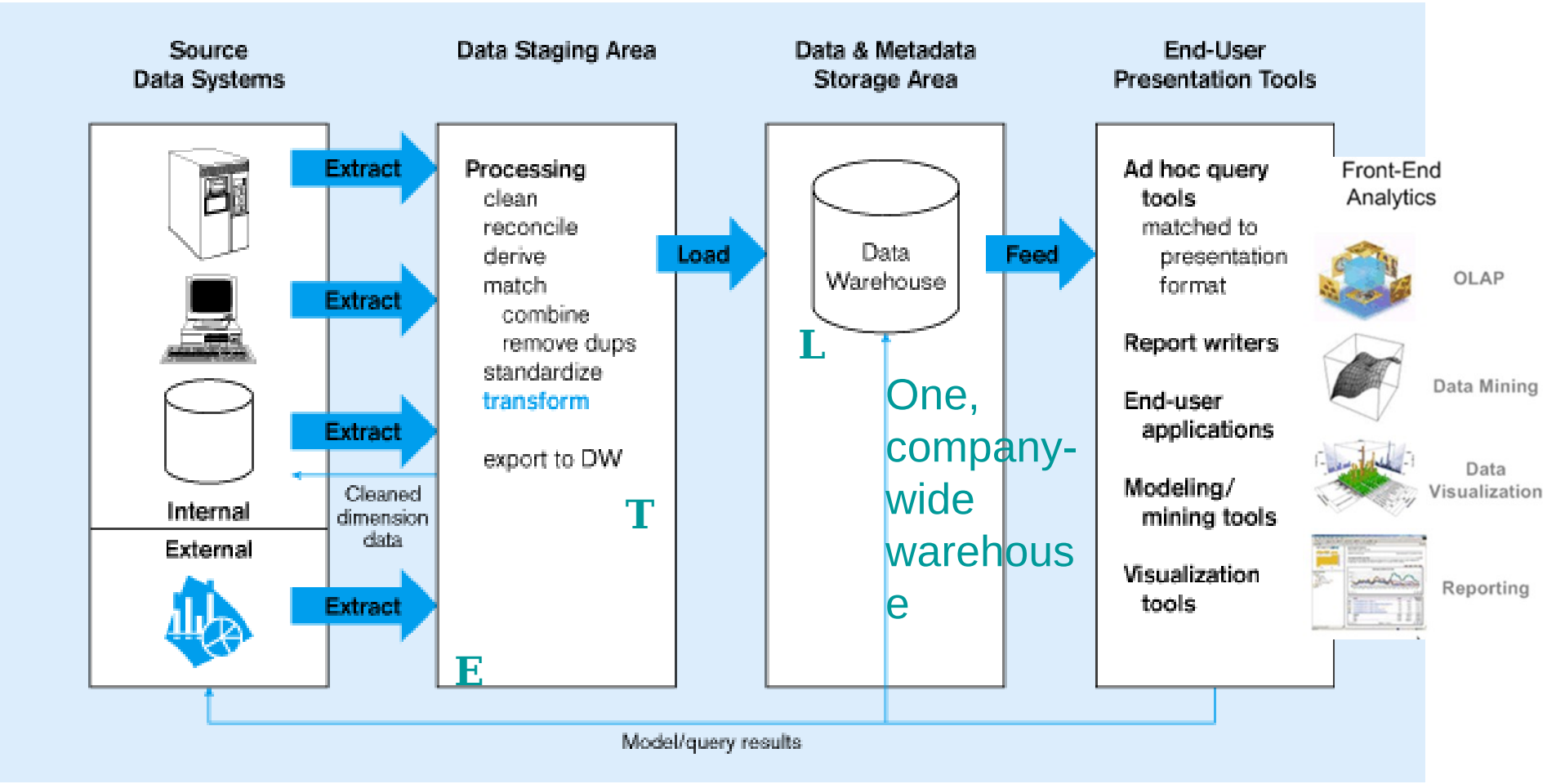
- A **data warehouse** is a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs.


Data Sources
Staging Area
Warehouse
Users
Operational System
Operational System
Flat Files
Metadata
Summary Data
Raw Data
Analysis
Reporting
Mining

Definition -2.2 (Gartner Group)

- The five components of a data warehouse are
  - production data sources,
  - data extraction and conversion,
  - the data warehouse database management system,
  - data warehouse administration and
  - business intelligence (BI) tools.

Figure 11-2: Generic two-level architecture



Periodic extraction ➔ data is not completely current in warehouse

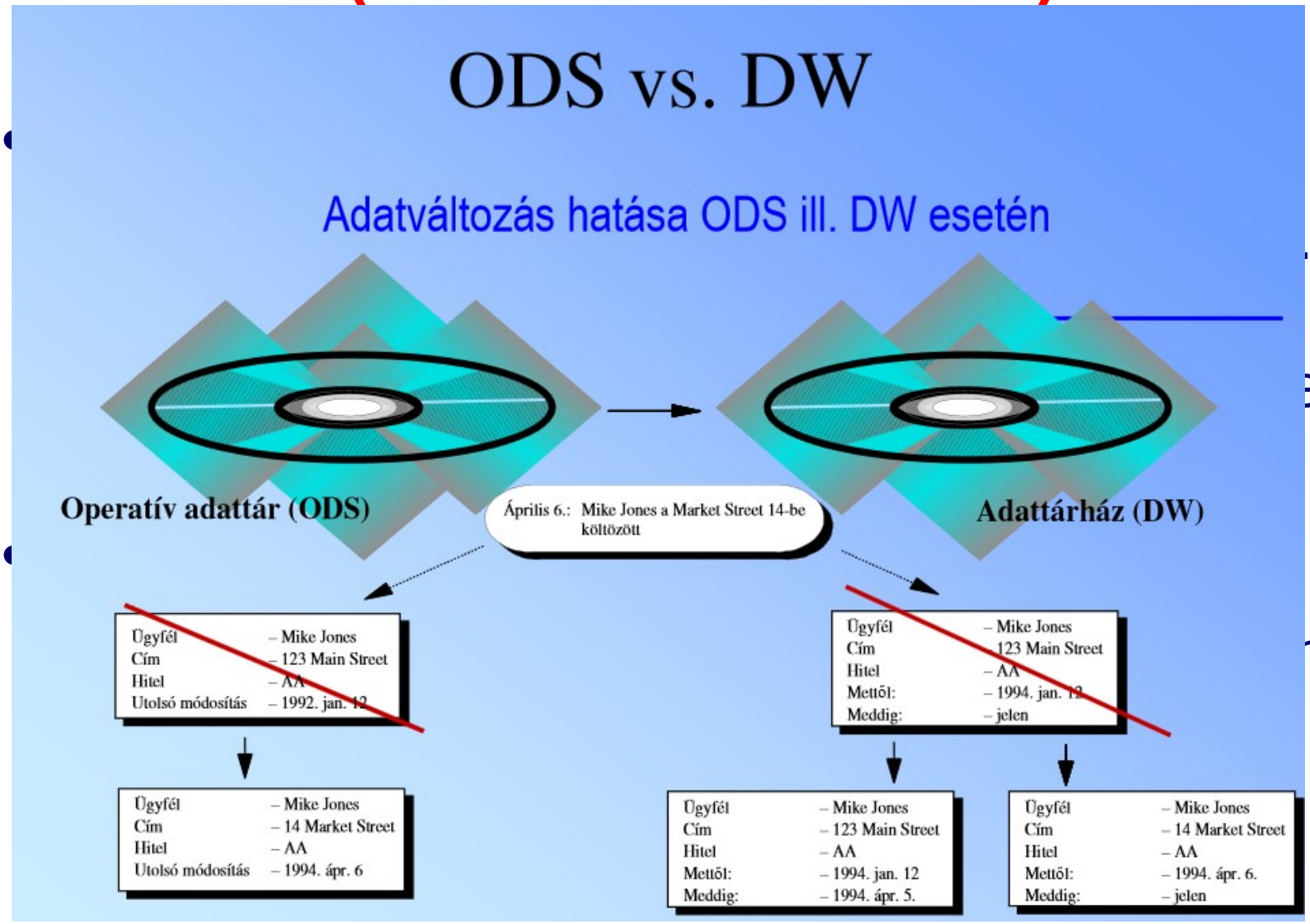
	Operatív adatkezelés	Döntéstámogató adatkezelés
Funkció	Adatfeldolgozás	Döntéstámogatás
Szállítandó	Üzleti műveletek támogatása	Információ
Szervezés	Folyamatorientált	Témaorientált
	Stabil, jól ismert folyamatokhoz	Ismeretlen, feltáró folyamatokhoz
Séma	Normalizált, sok tábla, kevés oszloppal	Kevesebb tábla, leíró oszlopok, redundancia
Lekérdezési profil	Ismert, ismétlődő lekérdezések	Aggregálások, sokféle lekérdezés
Érintett rekordszám	Kevés	Igen nagy
Válaszidő	Nagyon gyors	Lassabb
Frissítés	Folyamatos frissítés, sok tranzakció	Periodikus frissítás, bulk loading
Felhasználók száma	Nagy, egyidejűleg is	Jóval kevesebb
Számításigény	Stabil	Nagyon változó
Rendelkezésre állás	Kritikus, közel 100%	Nem kritikus

(Definition - 3)

(Bill Inmon)

- **Subject-Oriented**:
  - A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.
- **Integrated**:
  - A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

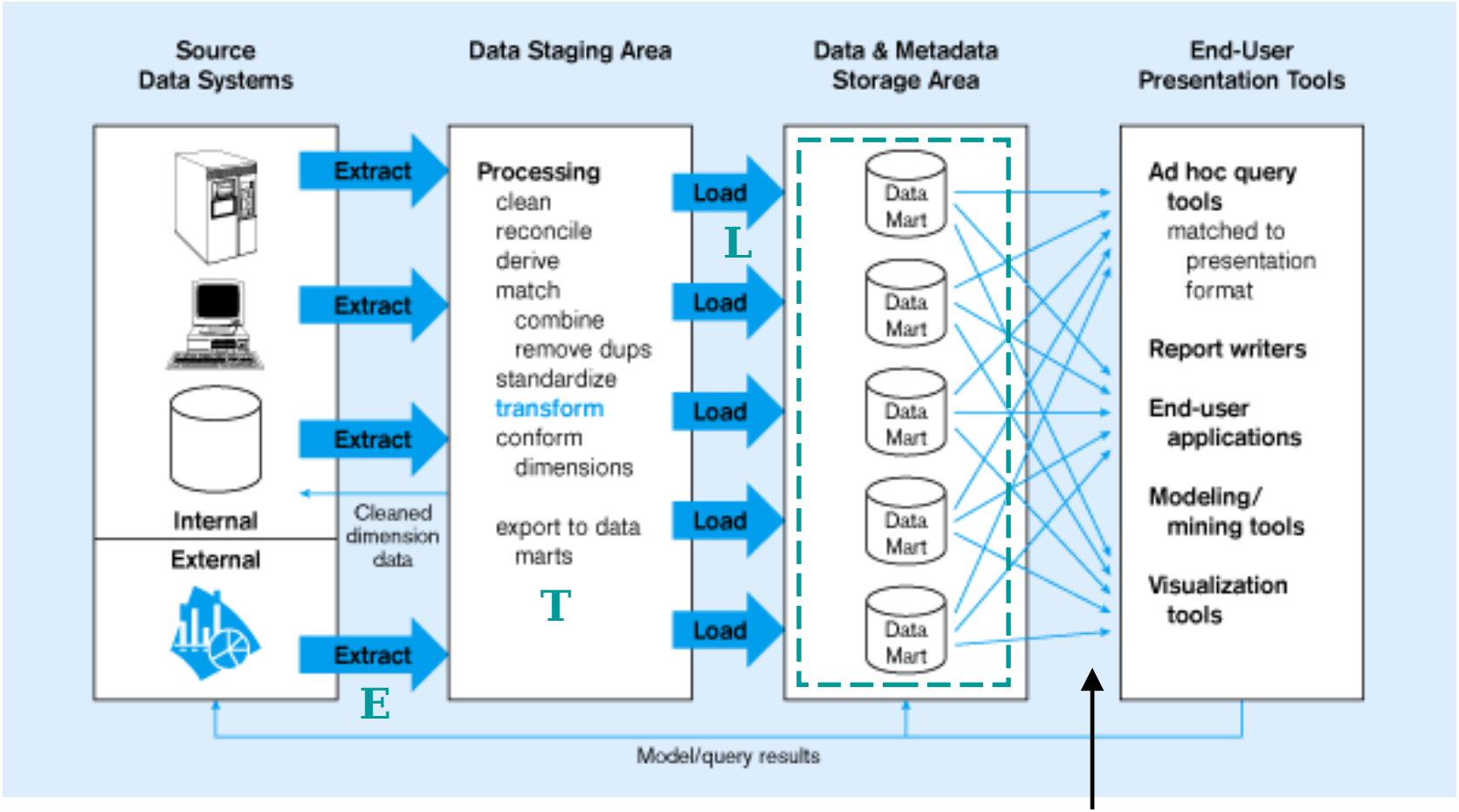
(Definition - 3.2)



(DATA MARTS)

Figure 11-3: Independent Data Mart

Data marts:
Mini-warehouses, limited in scope



Separate ETL for each ***independent*** data mart

Data access complexity due to ***multiple*** data marts

Table 11-2: Data Warehouse vs. Data Mart

Table 11-2 Data Warehouse Versus Data Mart

Data Warehouse	*Data Mart*
Scope	*Scope*
• Application independent	• Specific DSS application
• Centralized, possibly enterprise-wide	• Decentralized by user area
• Planned	• Organic, possibly not planned
Data	*Data*
• Historical, detailed, and summarized	• Some history, detailed, and summarized
• Lightly denormalized	• Highly denormalized
Subjects	*Subjects*
• Multiple subjects	• One central subject of concern to users
Sources	*Sources*
• Many internal and external sources	• Few internal and external sources
Other Characteristics	*Other Characteristics*
• Flexible	• Restrictive
• Data-oriented	• Project-oriented
• Long life	• Short life
• Large	• Start small, becomes large
• Single complex structure	• Multi, semi-complex structures, together complex

Adapted from Strange (1997)

Source: adapted from Strange (1997).

ETL PROCESSES, TOOLS

Data Reconciliation

- Typical operational data is:
  - Transient – not historical
  - Restricted in scope – not comprehensive
  - Sometimes poor quality – inconsistencies and errors
- After ETL, data should be:
  - Detailed – not summarized yet
  - Historical – periodic
  - Comprehensive – enterprise-wide perspective
  - Quality controlled – accurate with full integrity

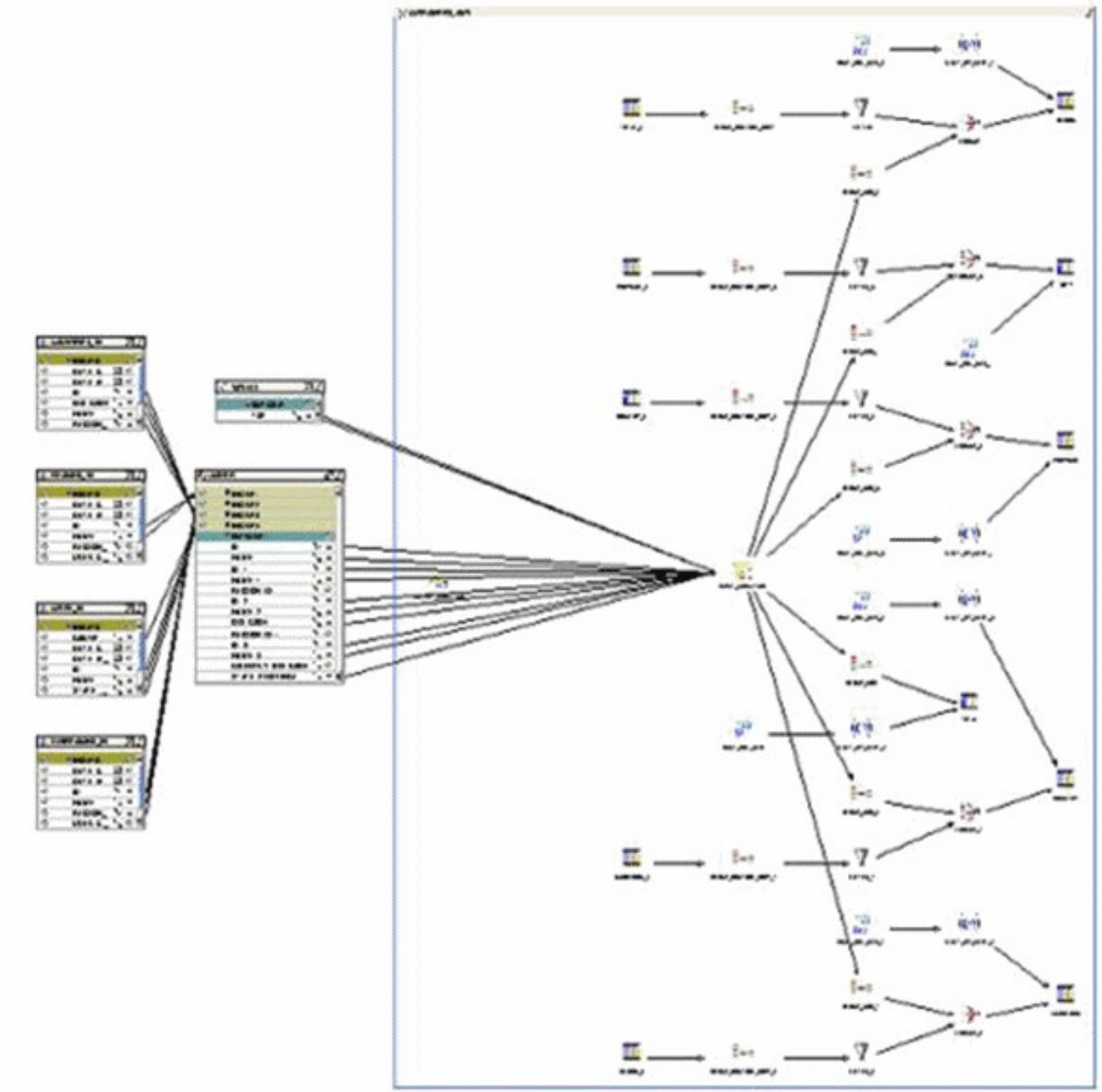
The ETL Process

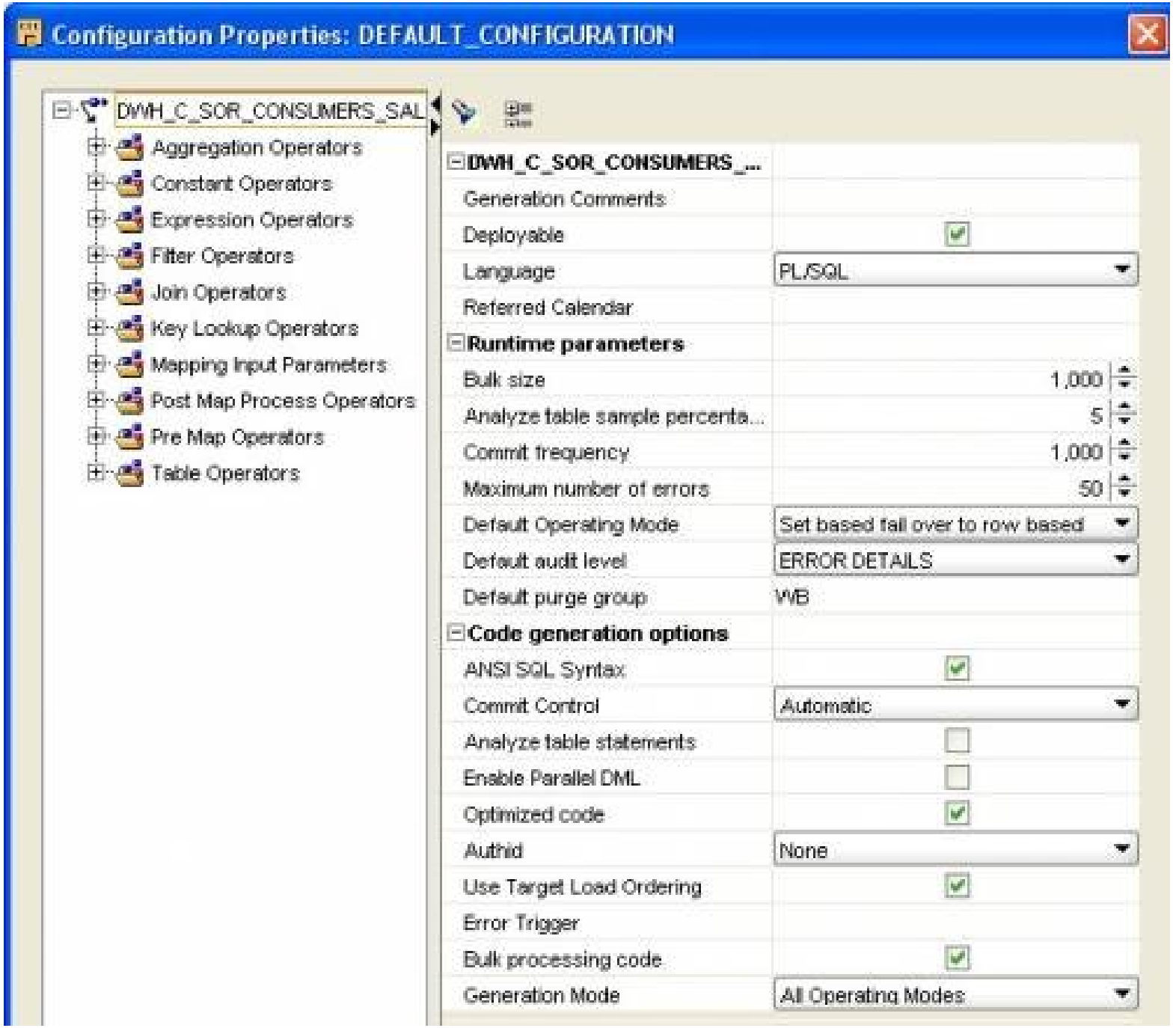
- Capture
- Scrub or data cleansing
- Transform
- Load and Index

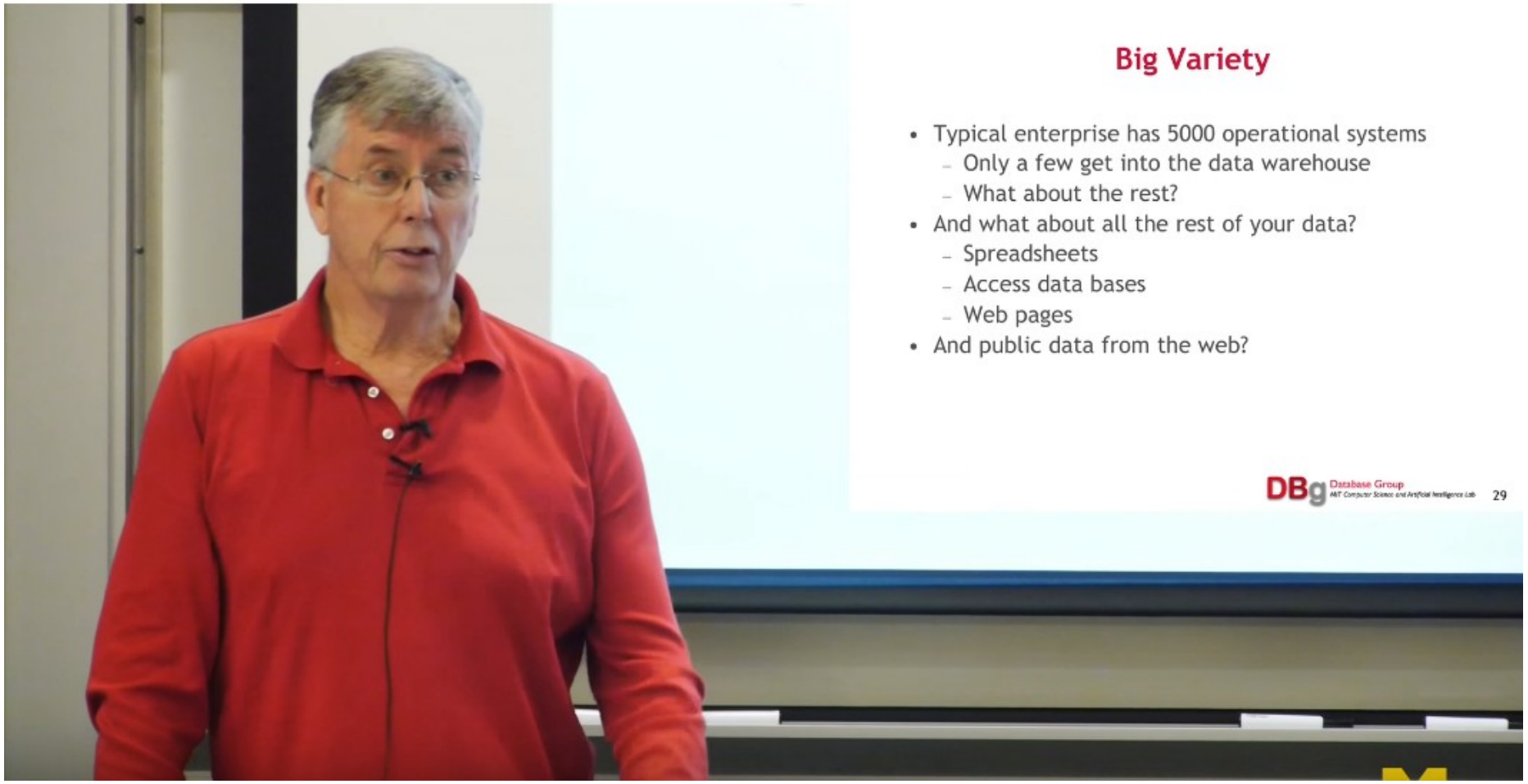
ETL = Extract, transform, and load

ETL Tools

- Declarative rules (what)
  - Mappings and transformations defined graphically
- Actual implementation (how)
  - Handling of errors
  - Performance
  - Auditing
  - Method
    - Also change data capture




Configuration Properties: DEFAULT_CONFIGURATION
DWH_C_SOR_CONSUMERS_SAL
Aggregation Operators
Constant Operators
Expression Operators
Filter Operators
Join Operators
Key Lookup Operators
Mapping Input Parameters
Post Map Process Operators
Pre Map Operators
Table Operators
DWH_C_SOR_CONSUMERS_...
Generation Comments
Deployable
Language
PL/SQL
Referred Calendar
Runtime parameters
Bulk size
1,000
Analyze table sample percenta...
5
Commit frequency
1,000
Maximum number of errors
50
Default Operating Mode
Set based fail over to row based
Default audit level
ERROR DETAILS
Default purge group
WB
Code generation options
ANSI SQL Syntax
Commit Control
Automatic
Analyze table statements
Enable Parallel DML
Optimized code
Authid
None
Use Target Load Ordering
Error Trigger
Bulk processing code
Generation Mode
All Operating Modes



https://youtu.be/KRcecxdGxvQ?t=2379

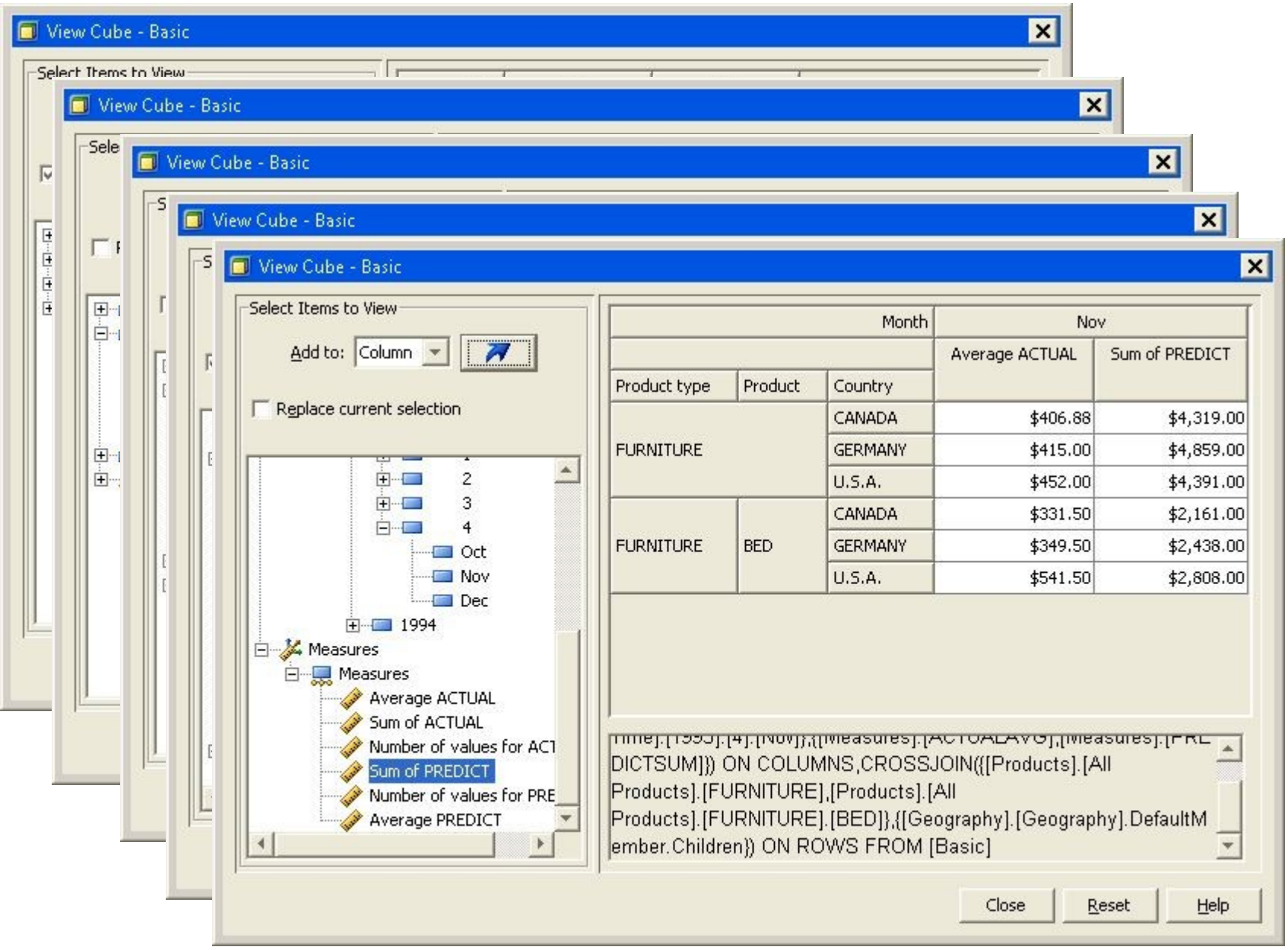
QUERYING TOOLS

(Querying Tools)

- SQL is not an analytical language
- SQL-99 includes some data warehousing extensions
- SQL-99 still is not a full-featured data warehouse querying and analysis tool.
- Different DBMS vendors will implement some or all of the SQL-99 OLAP extension commands and possibly others.

On-Line Analytical Processing (OLAP)

- OLAP is the use of a set of graphical tools that provides users with multidimensional views of their data and allows them to analyze the data using simple windowing techniques
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as *sales*, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as *item* (*item_name, brand, type*), or *time* (*day, week, month, quarter, year*)
  - Fact table contains measures (such as *dollars_sold*) and keys to each of the related dimension tables


View Cube - Basic
Select Items to View
Add to: Column
Replace current selection
2
3
4
Oct
Nov
Dec
1994
Measures
Measures
Average ACTUAL
Sum of ACTUAL
Number of values for ACT
Sum of PREDICT
Number of values for PRE
Average PREDICT
Month
Nov
Average ACTUAL
Sum of PREDICT
Product type
Product
Country
FURNITURE
CANADA
$406.88
$4,319.00
GERMANY
$415.00
$4,859.00
U.S.A.
$452.00
$4,391.00
FURNITURE
BED
CANADA
$331.50
$2,161.00
GERMANY
$349.50
$2,438.00
U.S.A.
$541.50
$2,808.00
DICTSUM]}) ON COLUMNS,CROSSJOIN({[Products].[All Products].[FURNITURE],[Products].[All Products].[FURNITURE].[BED]},{[Geography].[Geography].DefaultMember.Children}) ON ROWS FROM [Basic]
Close
Reset
Help

MOLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice:
  - *project and select*

Figure 11-22: Slicing a data cube

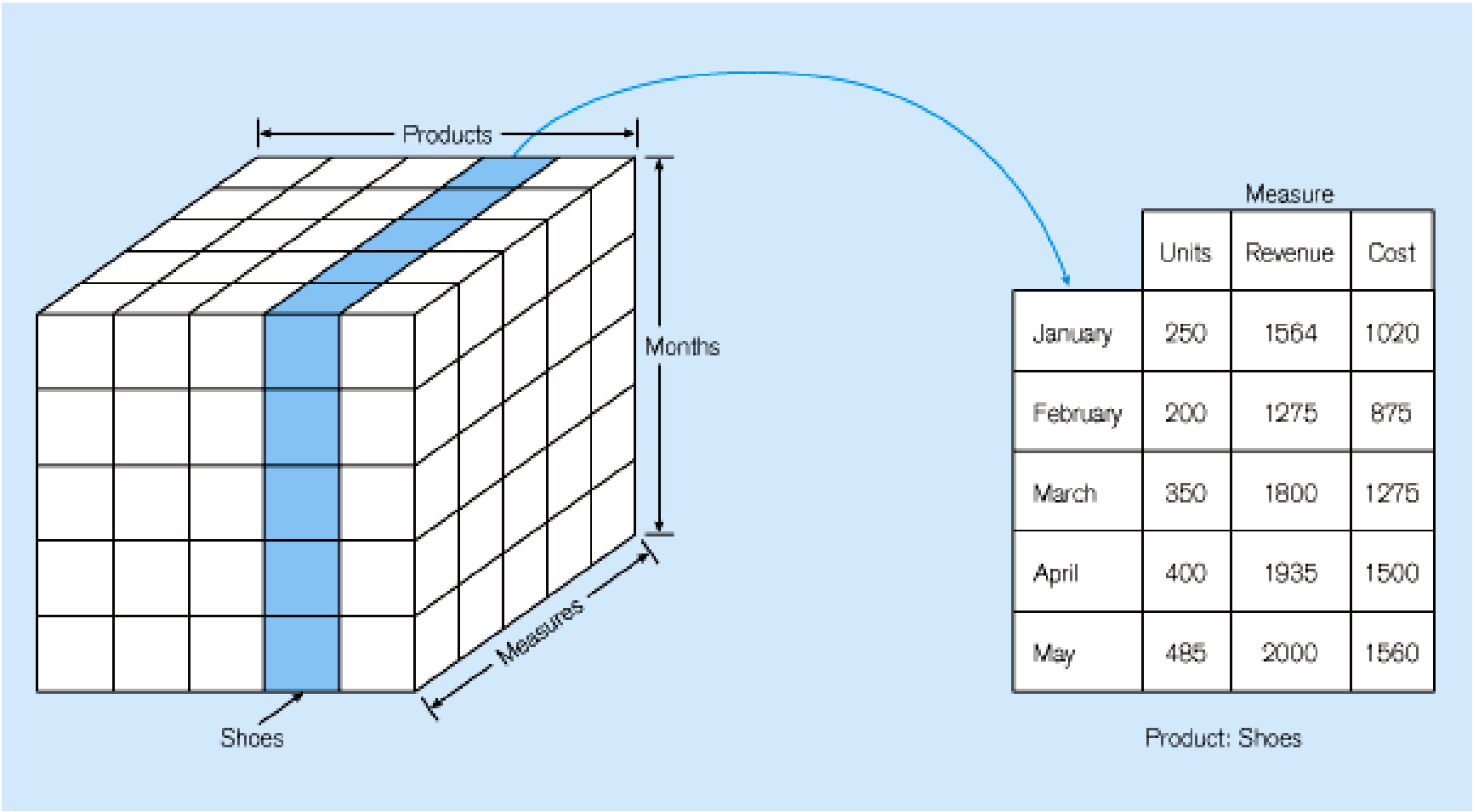


Figure 11-23:
Example of drill-down

Summary report

Brand	Package size	Sales
SofTowel	2-pack	$75
SofTowel	3-pack	$100
SofTowel	6-pack	$50

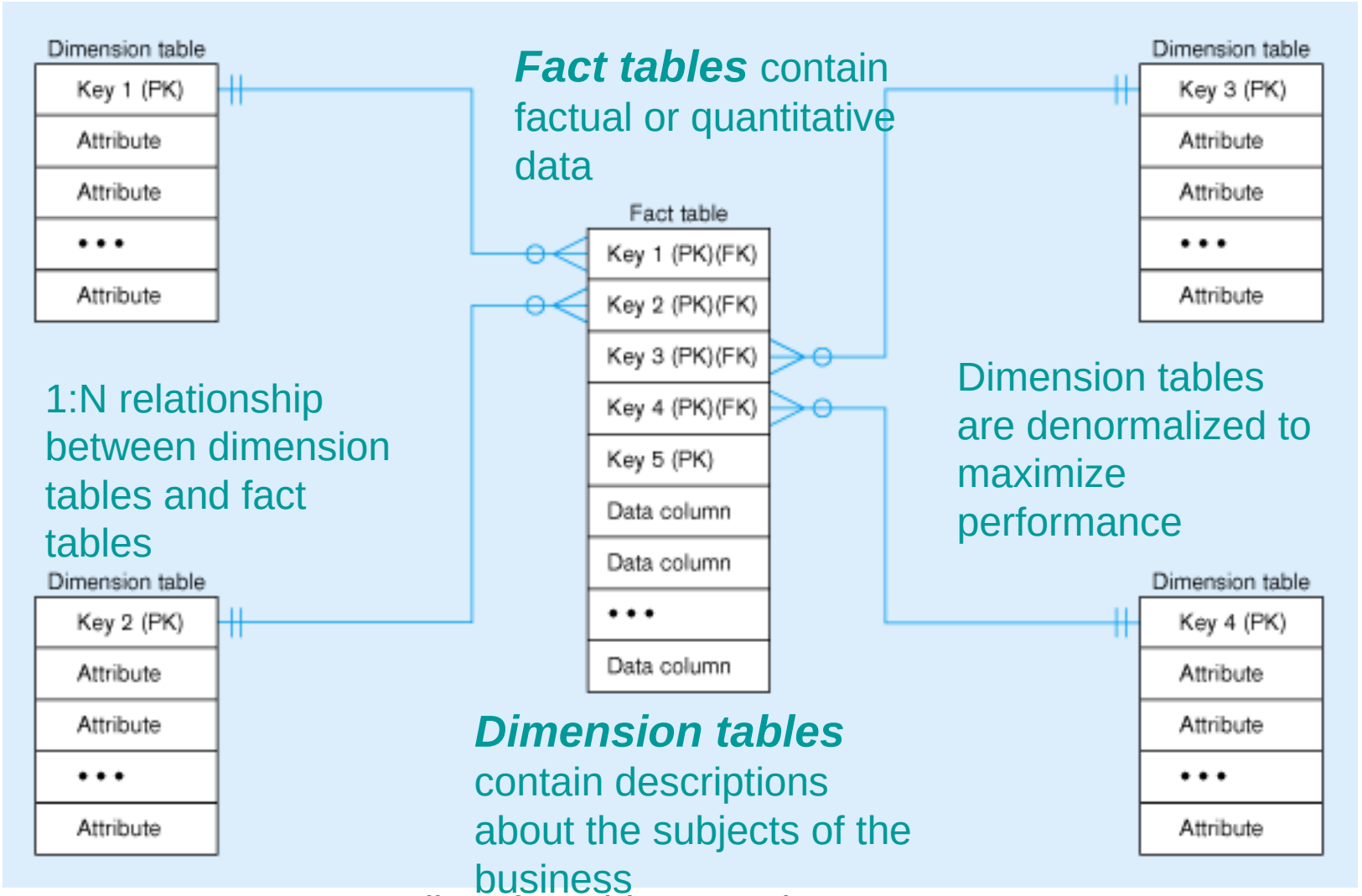
Drill-down with color added

Brand	Package size	Color	Sales
SofTowel	2-pack	White	$30
SofTowel	2-pack	Yellow	$25
SofTowel	2-pack	Pink	$20
SofTowel	3-pack	White	$50
SofTowel	3-pack	Green	$25
SofTowel	3-pack	Yellow	$25
SofTowel	6-pack	White	$30
SofTowel	6-pack	Yellow	$20

The Star Schema

- *Star schema*: is a simple database design in which dimensional (describing how data are commonly aggregated) are separated from fact or event data.
- A star schema consists of two types of tables: *fact tables* and *dimension table*.

Figure 11-13: Components of a **star schema**



Excellent for ad-hoc queries,
but bad for online transaction processing

Figure 11-14: Star schema example

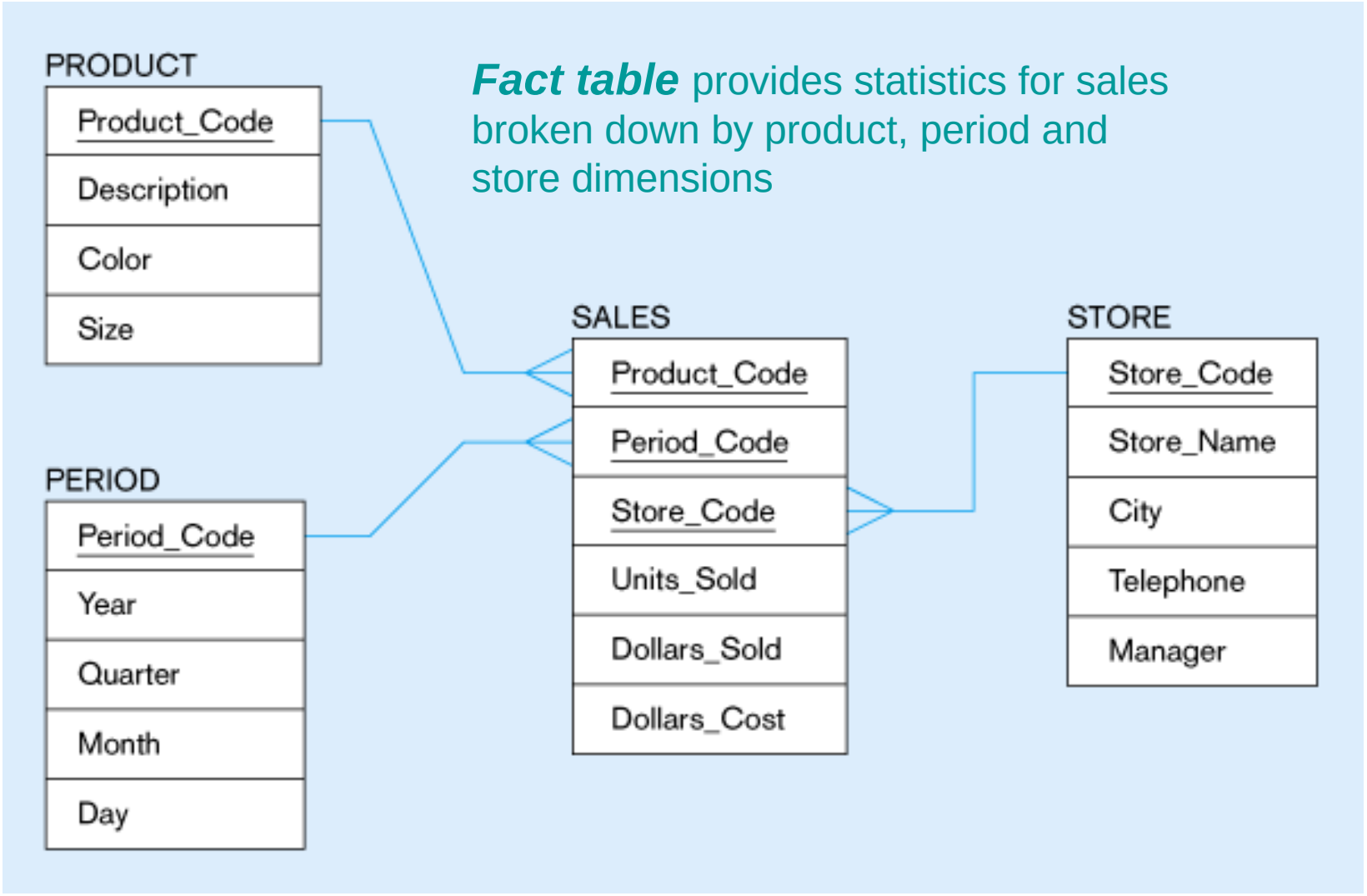
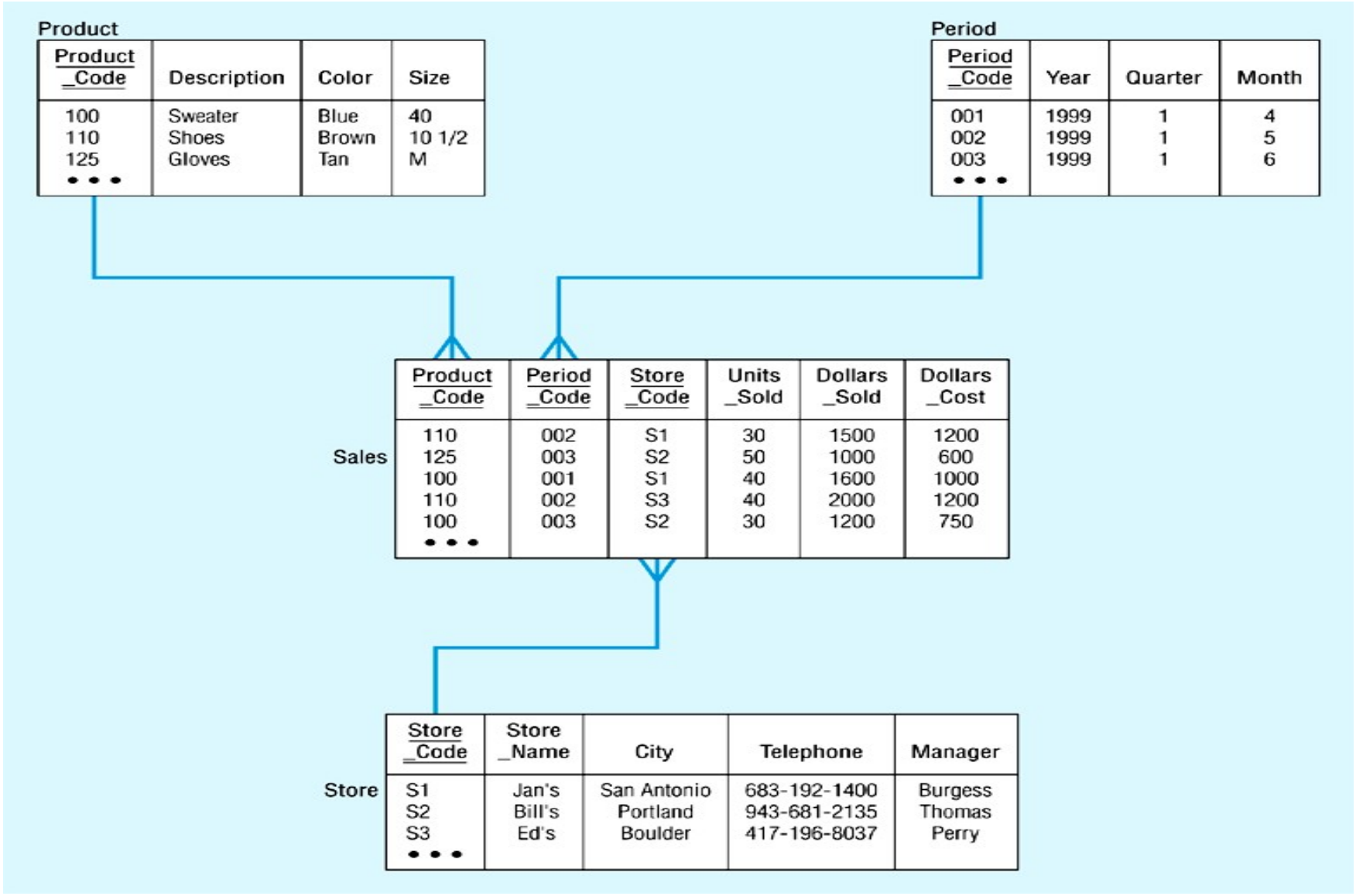


Figure 11-15: Star schema with sample data



(METADATA)

(Role of Metadata (data catalog))

- Identify subjects Identify dimensions and facts
- Indicate how data is derived from enterprise data warehouses, including derivation rules
- Indicate how data is derived from operational data store, including derivation rules
- Identify available reports and predefined queries
- Identify data analysis techniques (e.g. drill-down)
- Identify responsible people

DATA QUALITY

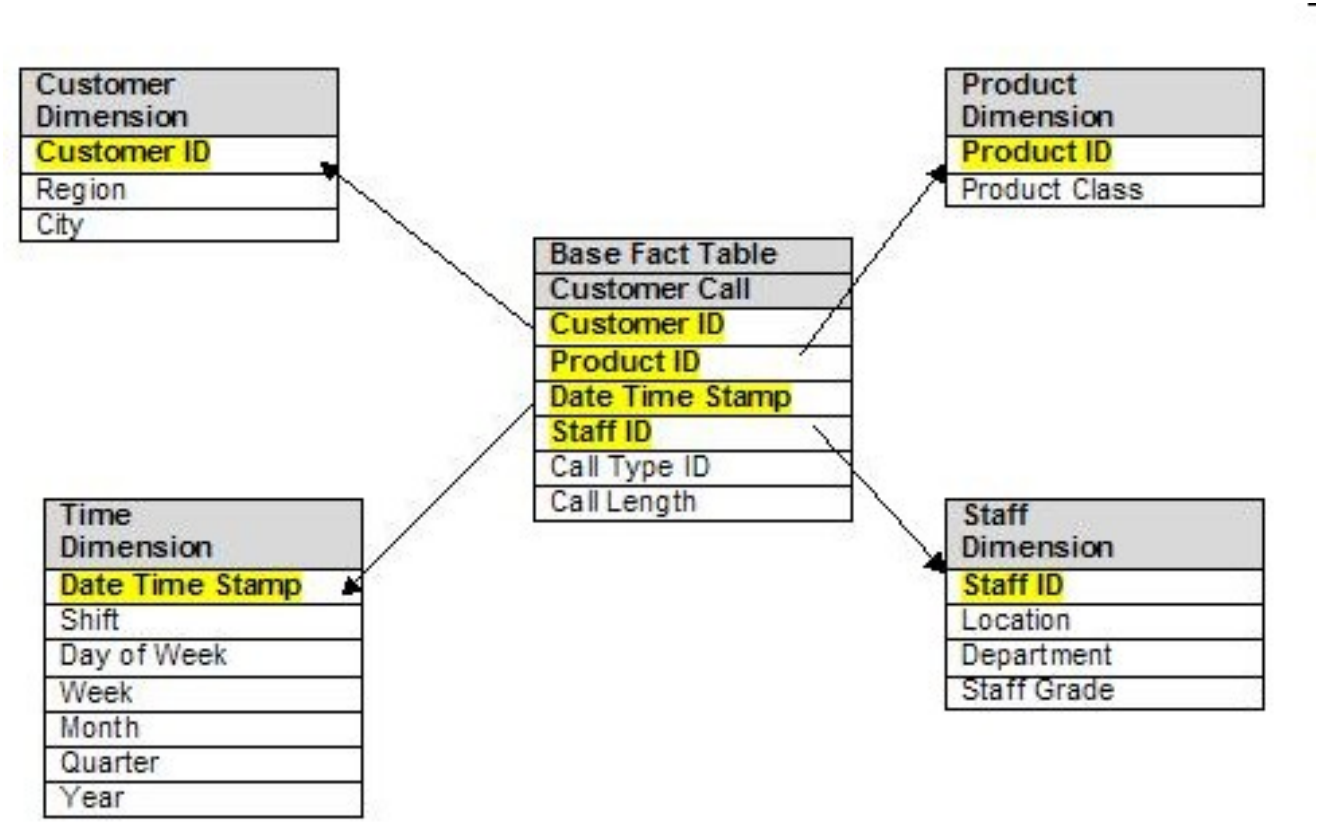
Optimisation

Data warehouse queries

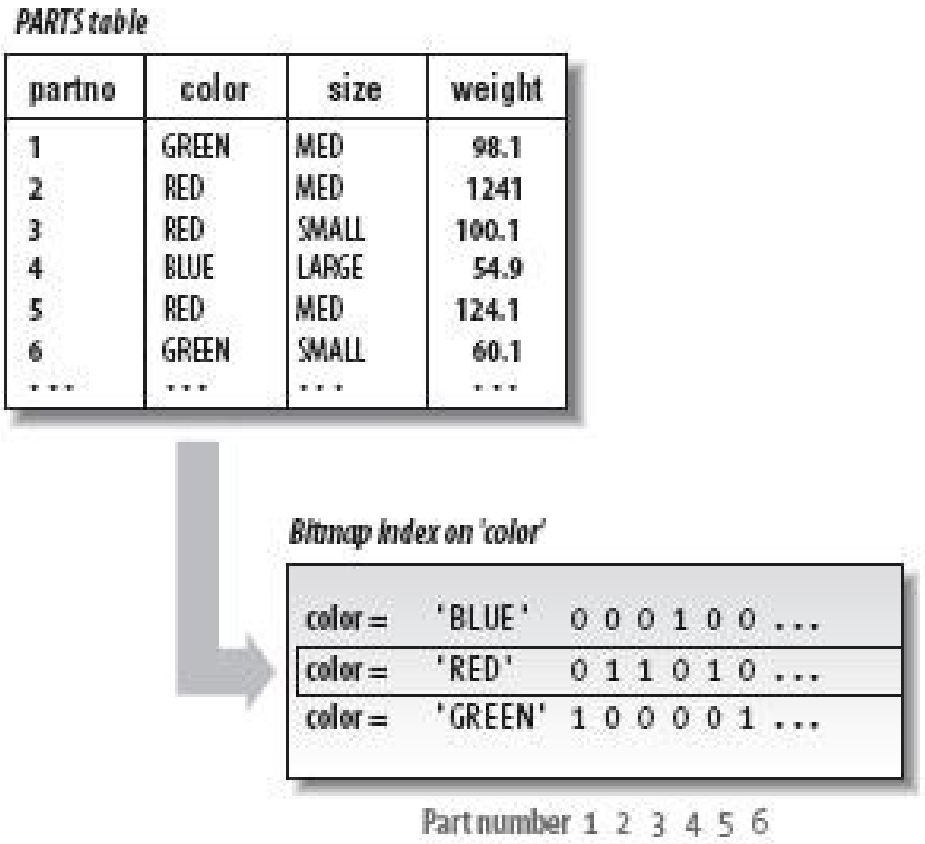
- Large number of records
- Ad-hoc queries, multiple dimensions
- Aggregations

- Read operations -

Star Schema (Bitmap index)



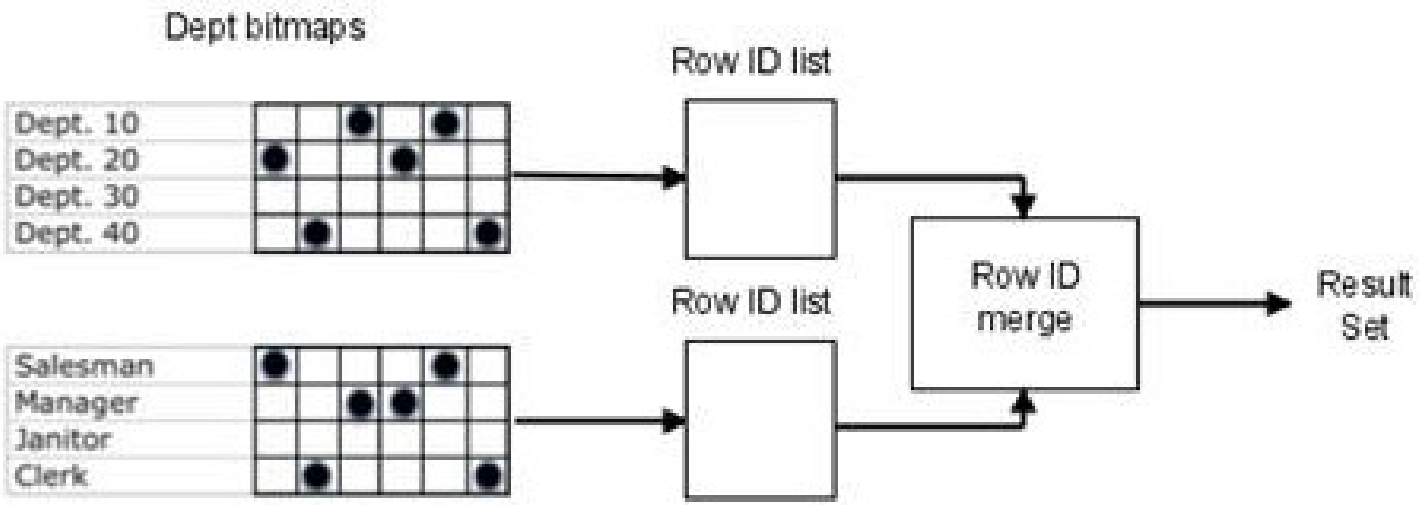
Bitmap index



+ compression

- **Low cardinality columns**
- **Infrequently updated or read-only tables**

Bitmap index



Bitmap index

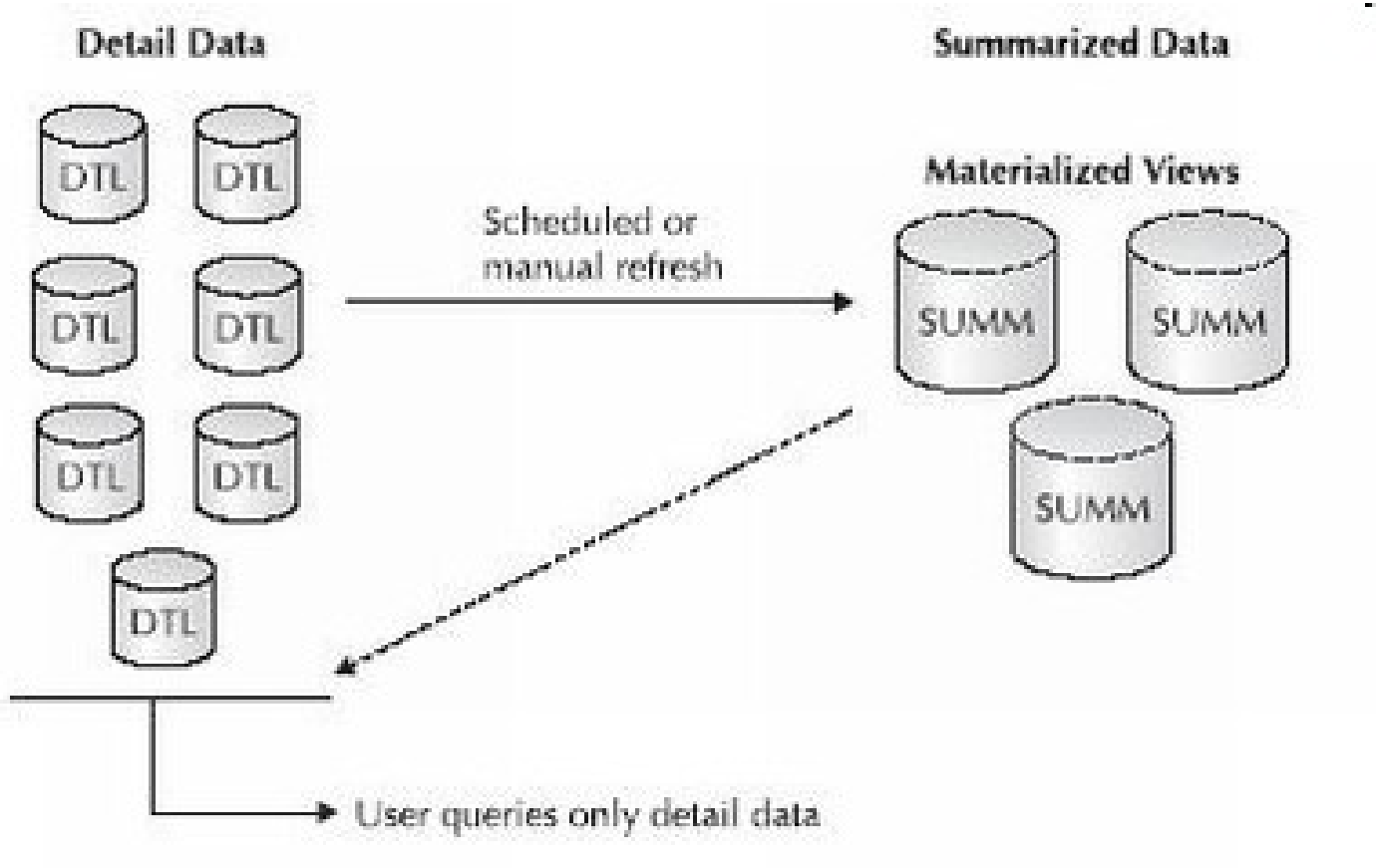
status = 'married'		region = 'central'		region = 'west'						
0		0		0		0		0		0
1		1		0		1		1		1
1	AND	0	OR	1	=	1	AND	1	=	1
0		0		1		0		1		0
0		1		0		0		1		0
1		1		0		1		1		1

Star transformation

```
SELECT ch.channel_class, c.cust_city, t.calendar_quarter_des
FROM sales s, times t, customers c, channels ch
WHERE s.time_id = t.time_id
AND s.cust_id = c.cust_id
AND s.channel_id = ch.channel_id
AND c.cust_state_province = 'CA'
AND ch.channel_desc IN ('Internet','Catalog')
AND t.calendar_quarter_desc IN (‘2006-Q1',‘2006-Q2')

SELECT ch.channel_class, c.cust_city, t.calendar_quarter_desc
FROM sales WHERE
time_id IN
  (SELECT time_id FROM times WHERE calendar_quarter_desc
      IN(‘2006-Q1',‘2006-Q2'))
AND cust_id IN
  (SELECT cust_id FROM customers WHERE cust_state_province='CA'
AND channel_id IN
  (SELECT channel_id FROM channels WHERE channel_desc IN
      ('Internet','Catalog'));
```

Materialized view, query rewrite

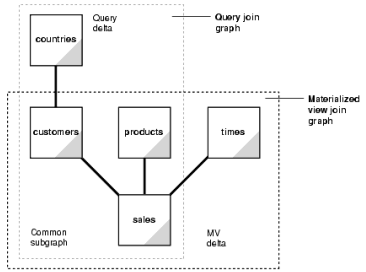


Materialized view

```
CREATE MATERIALIZED VIEW
  comm_prod_mv
  SELECT sales_rep_id, prod_id,
         comm_date, count(*),
         sum(comm_amt)
  FROM commission
  GROUP BY sales_rep_id,
           prod_id,  comm_date;
```

Materialized view, query rewrite

- Join Compatibility Check
  - Common joins that occur in both the query and the materialized view. These joins form the common subgraph.
  - Delta joins that occur in the query but not in the materialized view. These joins form the query delta subgraph.
  - Delta joins that occur in the materialized view but not in the query. These joins form the materialized view delta subgraph.
- Data Sufficiency Check
- Grouping Compatibility Check
- Aggregate Computability Check



Materialized view, query rewrite

Table 19-1 Dimension and Constraint Requirements for Query Rewrite

Query Rewrite Types	Dimensions	Primary Key/Foreign Key/Not Null Constraints
Matching SQL Text	Not Required	Not Required
Join Back	Required OR	Required
Aggregate Computability	Not Required	Not Required
Aggregate Rollup	Not Required	Not Required
Rollup Using a Dimension	Required	Not Required
Filtering the Data	Not Required	Not Required
PCT Rewrite	Not Required	Not Required
Multiple Materialized Views	Not Required	Not Required

Query AVG(X) MVI: SUM(X) and COUNT(

Query: years MVI: months

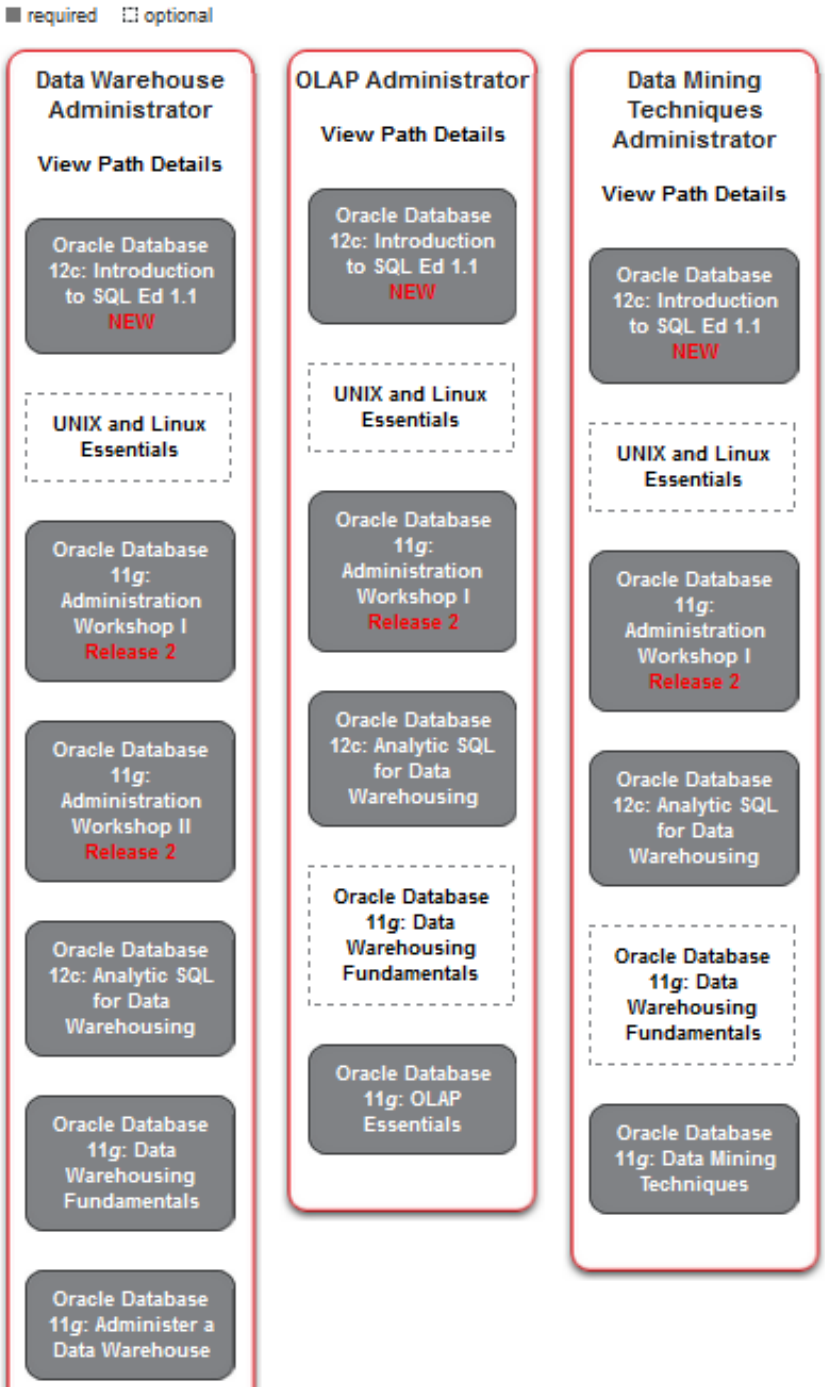
Refresh modes, query rewrite integrity

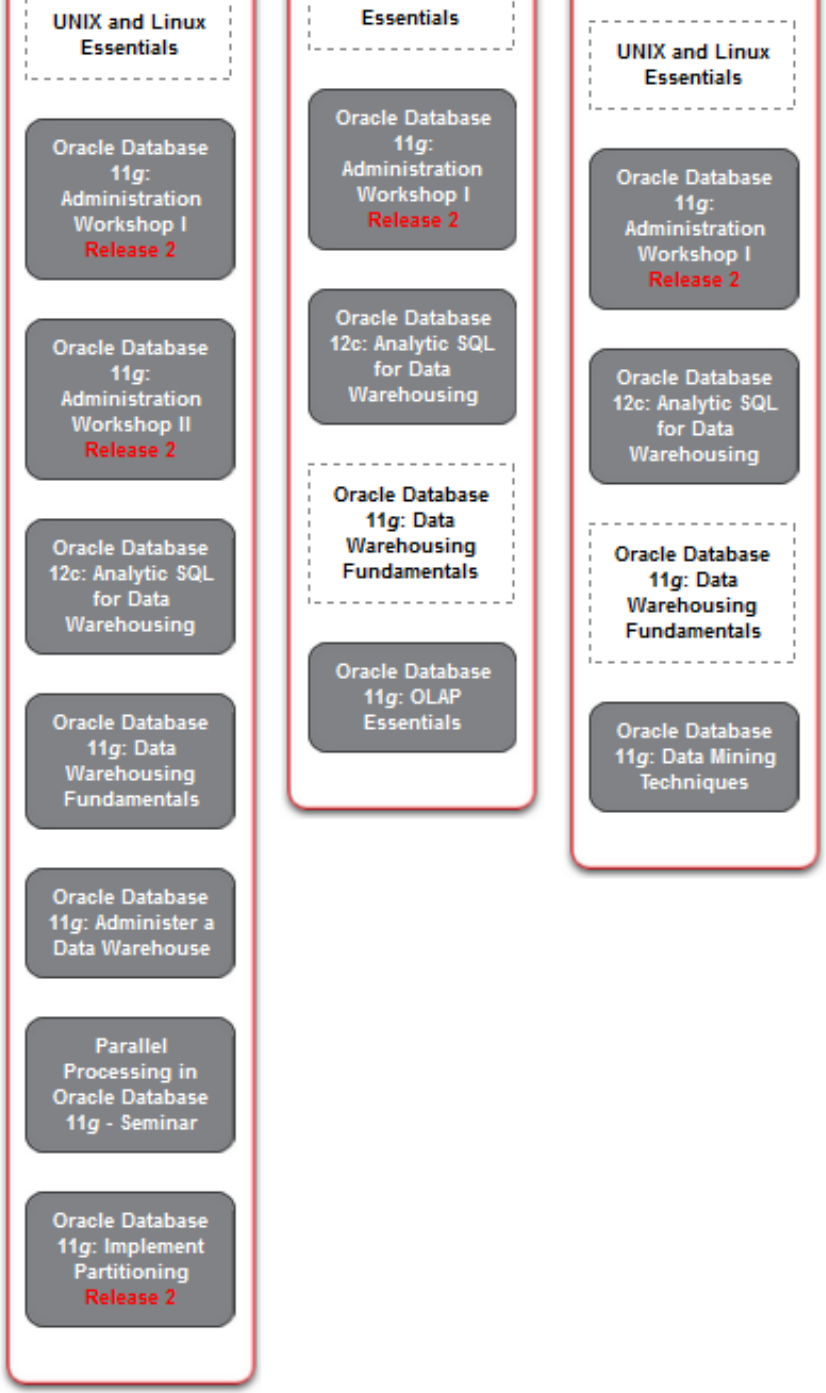
- Refresh types
  - Complete
  - Fast
- Accuracy of Query Rewrite (Oracle)
  - Enforced
  - Trusted
  - Stale_tolerated

További irodalom

- **Oracle® Database Data Warehousing Guide 11*g* Release 2 (11.2)**
  - http://docs.oracle.com/cd/E11882_01/server.112/e25554/toc.htm
  - **Basic Query Rewrite - http://docs.oracle.com/cd/E11882_01/server.112/e25554/qrbasic.htm#DWHSG018**
  - **Advanced Query Rewrite - http://docs.oracle.com/cd/E11882_01/server.112/e25554/qradv.htm#DWHSG080**
- **Oracle® Database 2 Day + Data Warehousing Guide 11*g* Release 2 (11.2)**
  - http://docs.oracle.com/cd/E11882_01/server.112/e25555/toc.htm
- **Data Warehousing and Business Intelligence**
  - http://docs.oracle.com/cd/E11882_01/nav/portal_6.htm

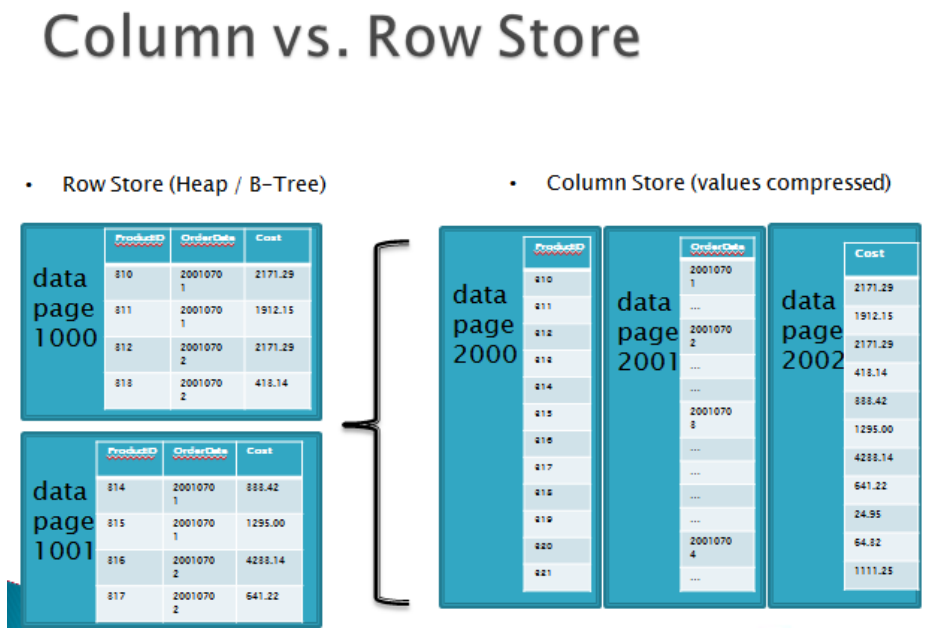
- Gajdos Sándor: Adattárház alapú vezet ői információs rendszerek https://db.bme.hu/~gajdos/2012adatb2/1.%20eloadas%20Analitikus%20alkalmazasok%20ppt.pdf
- Gajdos Sándor: Adattárházak és alkalmazásaik https://www.db.bme.hu/targyak/adattarhazak-alkalmazasaik

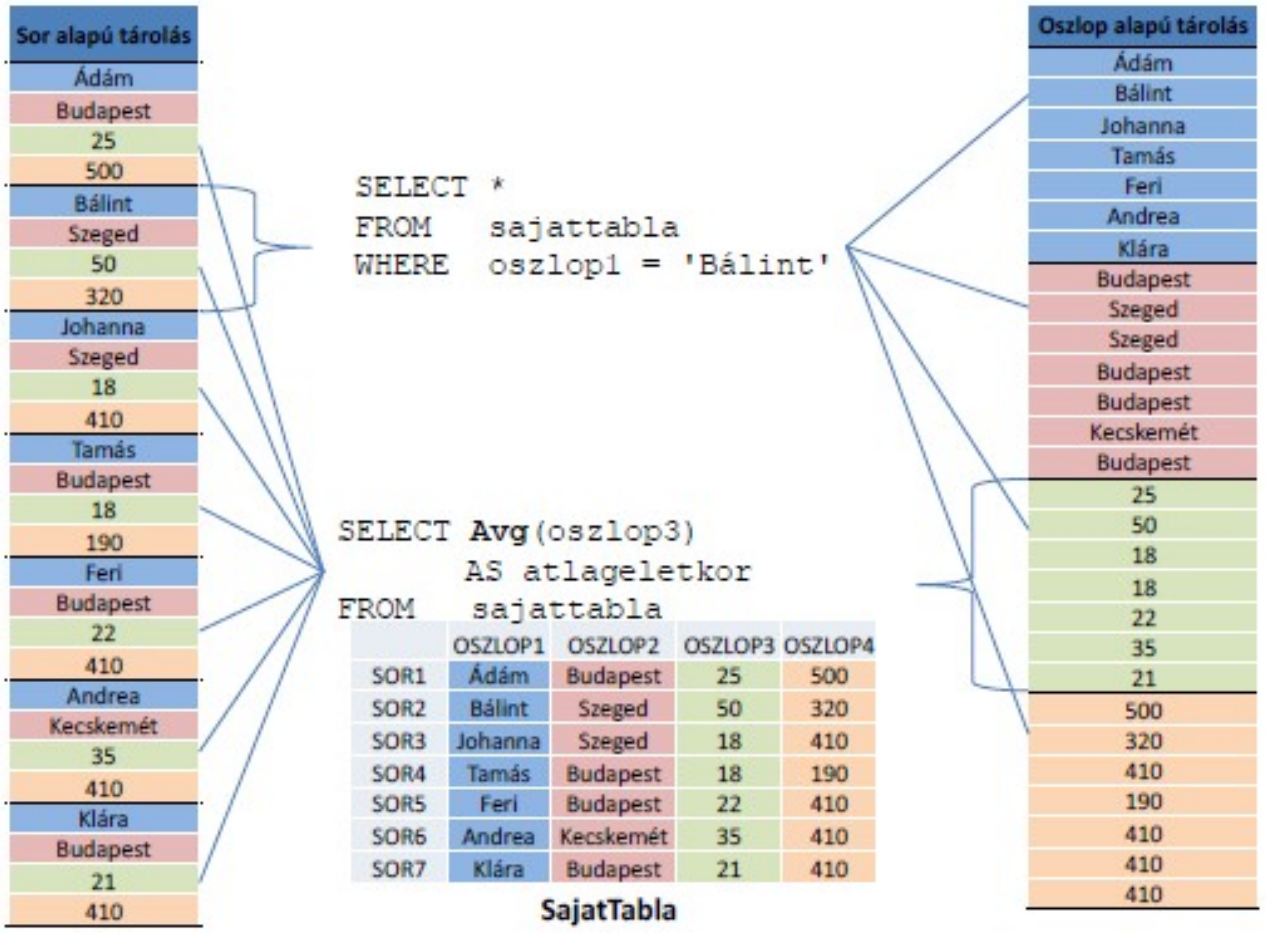

required
optional
Data Warehouse Administrator
View Path Details
Oracle Database 12c: Introduction to SQL Ed 1.1 NEW
UNIX and Linux Essentials
Oracle Database 11g: Administration Workshop I Release 2
Oracle Database 11g: Administration Workshop II Release 2
Oracle Database 12c: Analytic SQL for Data Warehousing
Oracle Database 11g: Data Warehousing Fundamentals
Oracle Database 11g: Administer a Data Warehouse
OLAP Administrator
View Path Details
Oracle Database 12c: Introduction to SQL Ed 1.1 NEW
UNIX and Linux Essentials
Oracle Database 11g: Administration Workshop I Release 2
Oracle Database 12c: Analytic SQL for Data Warehousing
Oracle Database 11g: Data Warehousing Fundamentals
Oracle Database 11g: OLAP Essentials
Data Mining Techniques Administrator
View Path Details
Oracle Database 12c: Introduction to SQL Ed 1.1 NEW
UNIX and Linux Essentials
Oracle Database 11g: Administration Workshop I Release 2
Oracle Database 12c: Analytic SQL for Data Warehousing
Oracle Database 11g: Data Warehousing Fundamentals
Oracle Database 11g: Data Mining Techniques


UNIX and Linux Essentials
Oracle Database 11g: Administration Workshop I Release 2
Oracle Database 11g: Administration Workshop II Release 2
Oracle Database 12c: Analytic SQL for Data Warehousing
Oracle Database 11g: Data Warehousing Fundamentals
Oracle Database 11g: Administer a Data Warehouse
Parallel Processing in Oracle Database 11g - Seminar
Oracle Database 11g: Implement Partitioning Release 2
Essentials
Oracle Database 11g: Administration Workshop I Release 2
Oracle Database 12c: Analytic SQL for Data Warehousing
Oracle Database 11g: Data Warehousing Fundamentals
Oracle Database 11g: OLAP Essentials
UNIX and Linux Essentials
Oracle Database 11g: Administration Workshop I Release 2
Oracle Database 12c: Analytic SQL for Data Warehousing
Oracle Database 11g: Data Warehousing Fundamentals
Oracle Database 11g: Data Mining Techniques

Column-oriented DBMS

- Column-oriented
  - Aggregates over many rows and few columns
  - Compression…
- Row-oriented
  - Retrieving or changing few records with many attributes





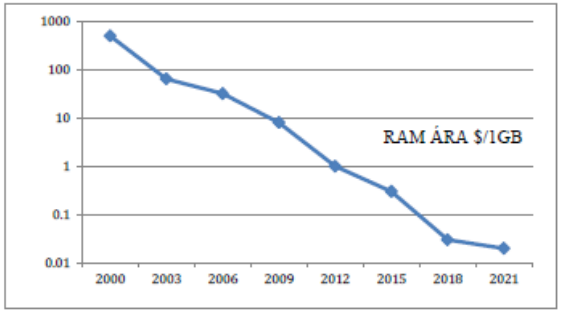
	OSZLOP1	OSZLOP2	OSZLOP3	OSZLOP4
SOR1	Ádám	Budapest	25	500
SOR2	Bálint	Szeged	50	320
SOR3	Johanna	Szeged	18	410
SOR4	Tamás	Budapest	18	190
SOR5	Feri	Budapest	22	410
SOR6	Andrea	Kecskemét	35	410
SOR7	Klára	Budapest	21	410

(Sípos Zsófia, szakdolgozat)

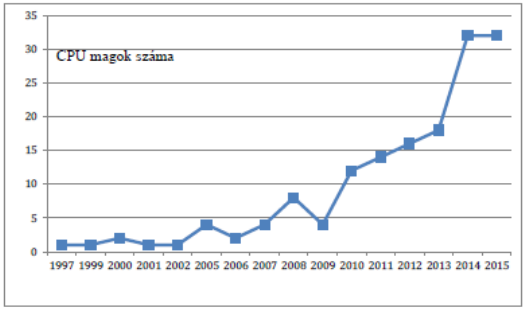
In-memory databases

(részben Sípos Zsófi szakdolgozatából)

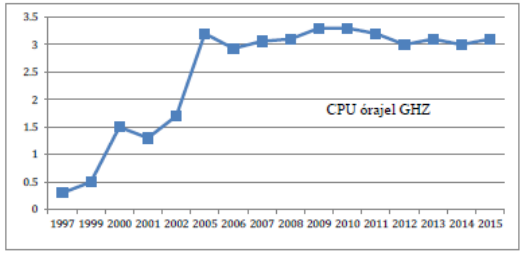
- RAM: szerver 1-6 TB !
- CPU magok száma, frekvencia
- SSD
  - (Logfájlok is!)



3. ábra A RAM árának csökkenése az elmúlt és a közeljövő években $/1GB-ban



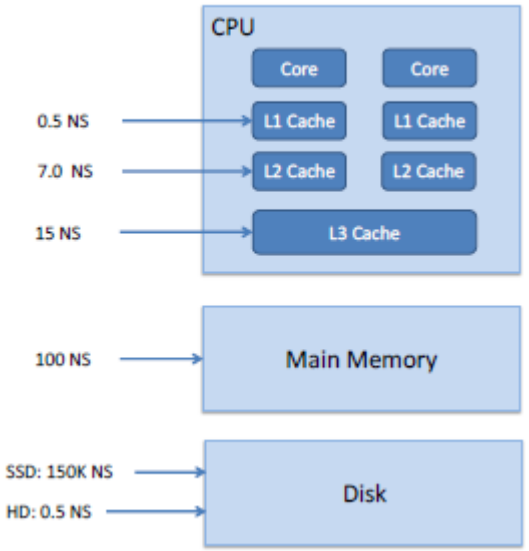
1. ábra a CPU magjainak számának növekedése az elmúlt években

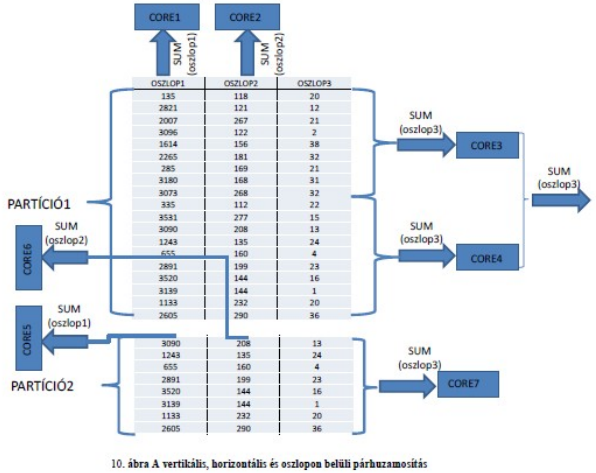


2. ábra a CPU órajelének növekedése az elmúlt években GHZ-ben

SAP HANA (High Speed Analytic Appliance)

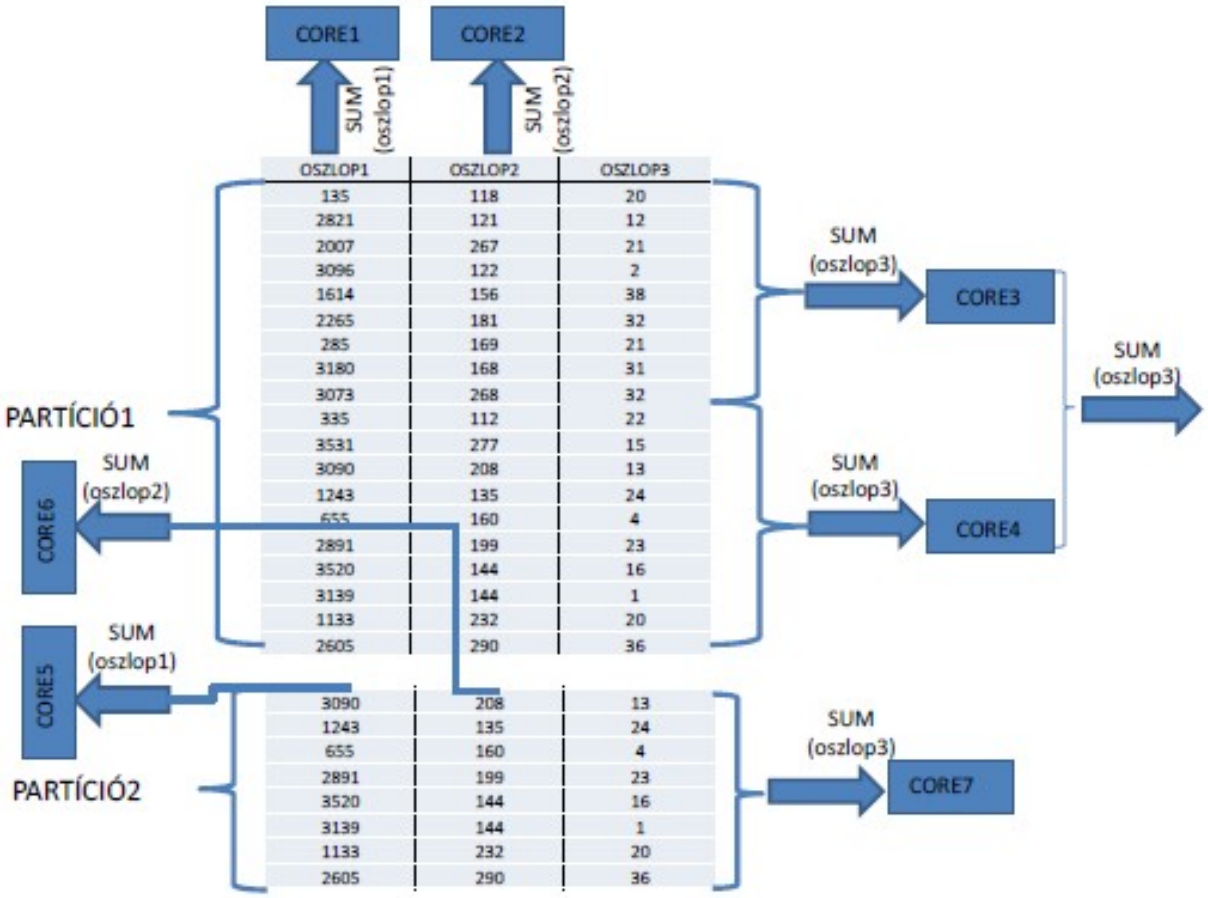
- Oszlop alapú tárolás is
  - Memóriából is gyorsabb szekvenciálisan olvasni
  - Tömörítés
- Párhuzamosítás





10. ábra A vertikális, horizontális és oszlopon belüli párhuzamosítás

SAP HANA
(High Speed Analytic Appliance)



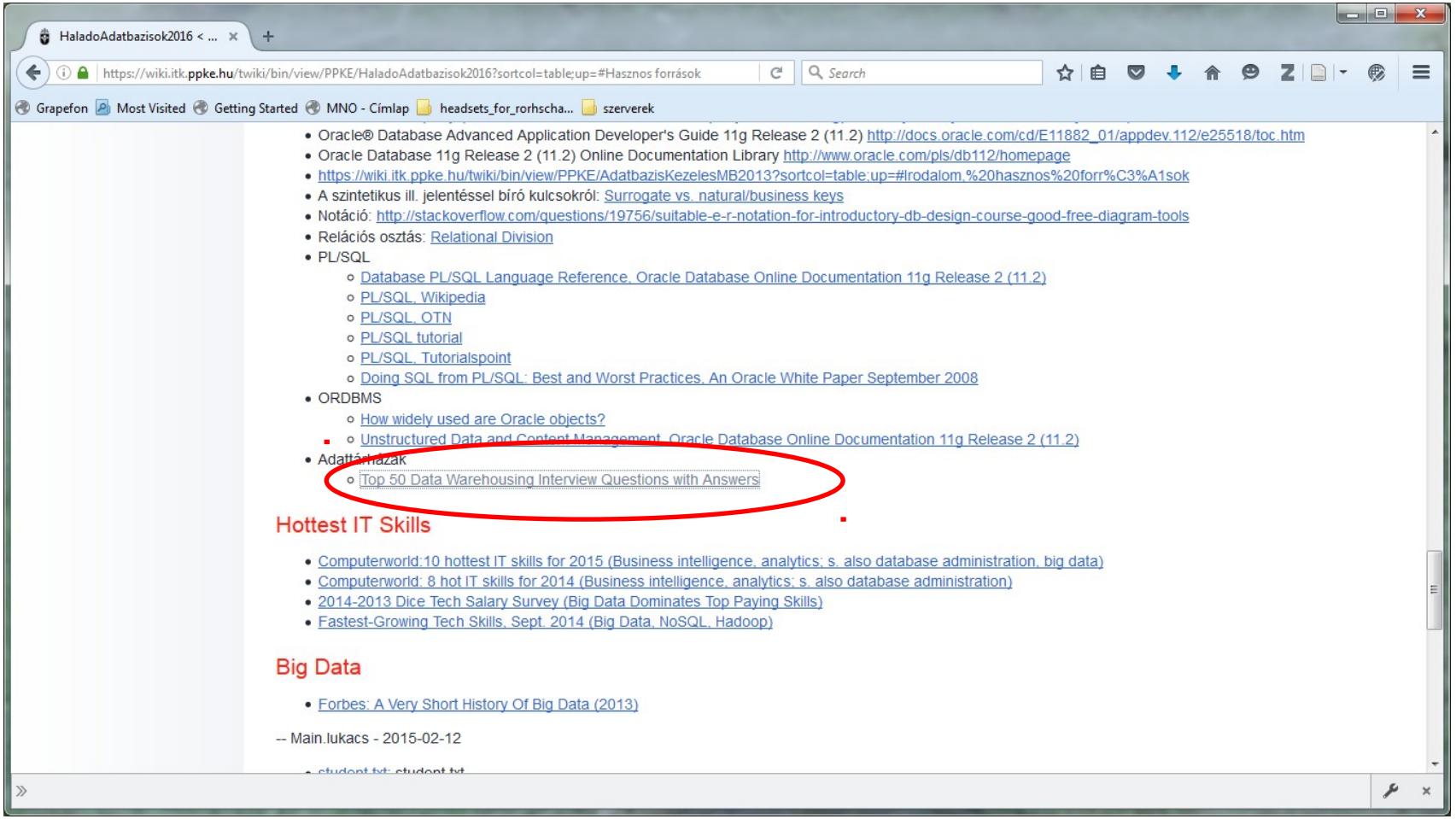
10. ábra A vertikális, horizontális és oszlopon belüli párhuzamosítás

In-memory adatbáziskezelés, adattárház?

- Kb. 2 nagyságrenddel gyorsabb adatfeldolgozás (pl. 6 óra -> 6 perc)
- OLTP és OLAP együtt!

- Oracle Database In-Memory
  - http://www.oracle.com/technetwork/database/in-memory/overview/twp-oracle-database-in-memory-2245633.html

Top 50 Data Warehousing Interview Questions with Answers



A következő materializált nézetek vannak **elkészítve** az adatbázisban (minden a ad18___db sémában) -- csak tájékoztatásul:

```
CREATE materialized VIEW historyitems_large_mv2 ENABLE QUERY REWRITE
AS
  SELECT user_id,
    SUM(duration),
    COUNT(duration)
  FROM historyitems_large
  GROUP BY user_id;


CREATE materialized VIEW historyitems_large_mv3 ENABLE QUERY REWRITE
AS
  SELECT user_id,
    TO_CHAR(started_at, 'YYYY-MM'),
    SUM(duration),
    COUNT(duration)
  FROM lukacs.historyitems_large
  GROUP BY user_id,
    TO_CHAR(started_at, 'YYYY-MM');


CREATE materialized VIEW historyitems_large_mv4 ENABLE QUERY REWRITE
AS
  SELECT user_id,
    TO_CHAR(started_at, 'YYYY'),
    SUM(duration),
    COUNT(duration)
  FROM ad18___db.historyitems_large
  GROUP BY user_id,
    TO_CHAR(started_at, 'YYYY');
```

Nézzük meg a következő lekérdezések lekérdezési tervét és lekérdezési költségét így és a lekérdezés újraírás letiltásával (/*+ NOREWRITE */ hint)

```
SELECT user_id, AVG(duration) FROM ad18___db.historyitems_large GROUP BY user_id;

SELECT DISTINCT user_id FROM ad18___db.historyitems_large;

SELECT SUBSTR(users.surname,8,1),
  SUM(historyitems_large.duration)
FROM ad18___db.historyitems_large
LEFT OUTER JOIN ad18___db.users
ON historyitems_large.user_id = users.id
GROUP BY SUBSTR(users.surname,8,1)
ORDER BY SUM(historyitems_large.duration) DESC;

SELECT user_id,
  TO_CHAR(started_at, 'YYYY'),
  AVG(duration)
FROM ad18___db.historyitems_large
GROUP BY user_id,
  TO_CHAR(started_at, 'YYYY');
```

A Query rewrite mechanizmus meglehetősen bonyolult, sokszor nehéz látni, miért nem sikerült az Oracle-nek a lekérdezést újraírnia.
Ezért biztosít egy eljárást, aminek egy lekérdezést és egy materializált nézet nevet string-ként átadva információt ad a lekérdezés újraírásról, ill. az újraírás akadályáról. Az információkat egy fix sémájú, rewrite_table nevű táblába írja.

Sajnos jogosultság hiányában nem tudják maguk végrehajtani, de itt egy példa végrehajtás:

```
BEGIN
     DBMS_MVIEW.EXPLAIN_REWRITE('select historyitems_large.user_id, sum(historyitems_large.duration),
sum(audio_large.duration)
from historyitems_large left outer join AUDIO_LARGE on historyitems_large.audio_id = audio_large.id
group by historyitems_large.user_id','HISTORYITEMS_LARGE_MV2','100');
end;
/

select message from rewrite_table;

QSM-01150: a lekérdezés újraírása nem sikerült
QSM-01067: a(z) HISTORYITEMS_LARGE_MV2 statikus nézet nem támogatja a(z) SUM lekérdezés-mérőszámot
QSM-01082: A(z) HISTORYITEMS_LARGE_MV2 statikus nézet összekapcsolása a(z) HISTORYITEMS_LARGE táblával
nem lehetséges
QSM-01102: a(z) HISTORYITEMS_LARGE_MV2 statikus nézethez visszakapcsolás szükséges a(z)
HISTORYITEMS_LARGE táblához a(z) AUDIO_ID oszlopban
QSM-01219: nem található megfelelő statikus nézet a lekérdezés újraírásához
```