Data Warehouse

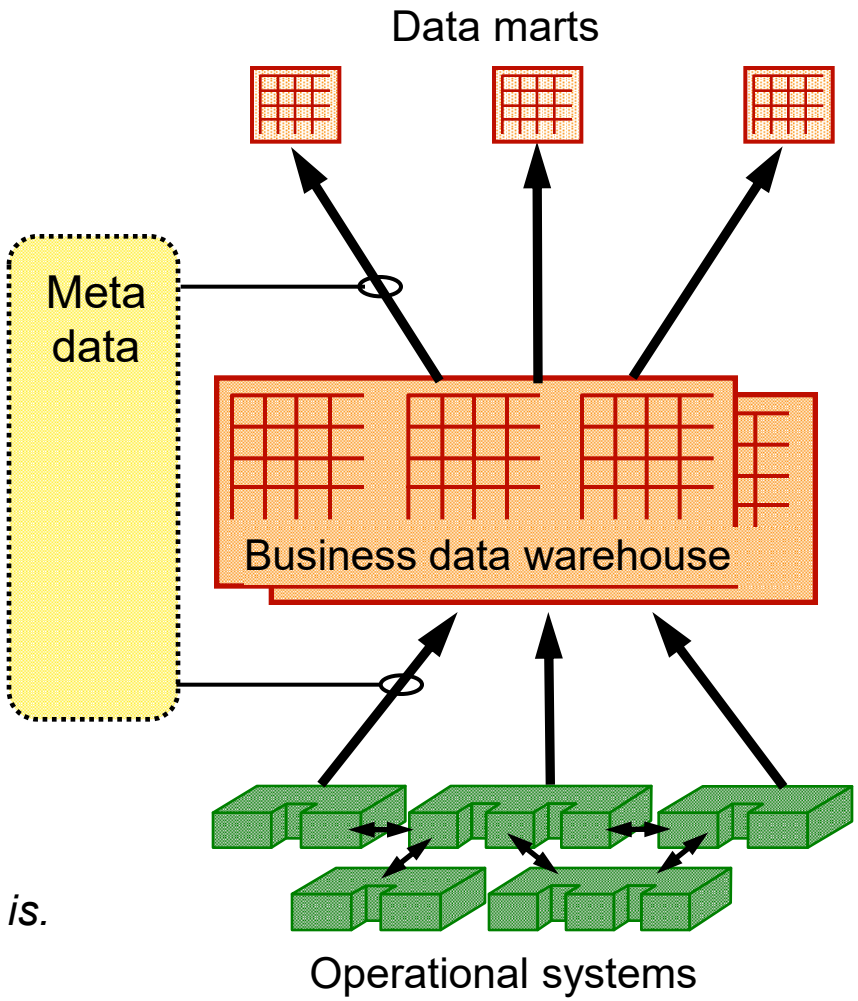
Martin Clement
Teamleiter Analytics Consulting
Martin.Clement@atvantage.com

Data Warehouse History

Mid/Late -1980s: Data Warehousing

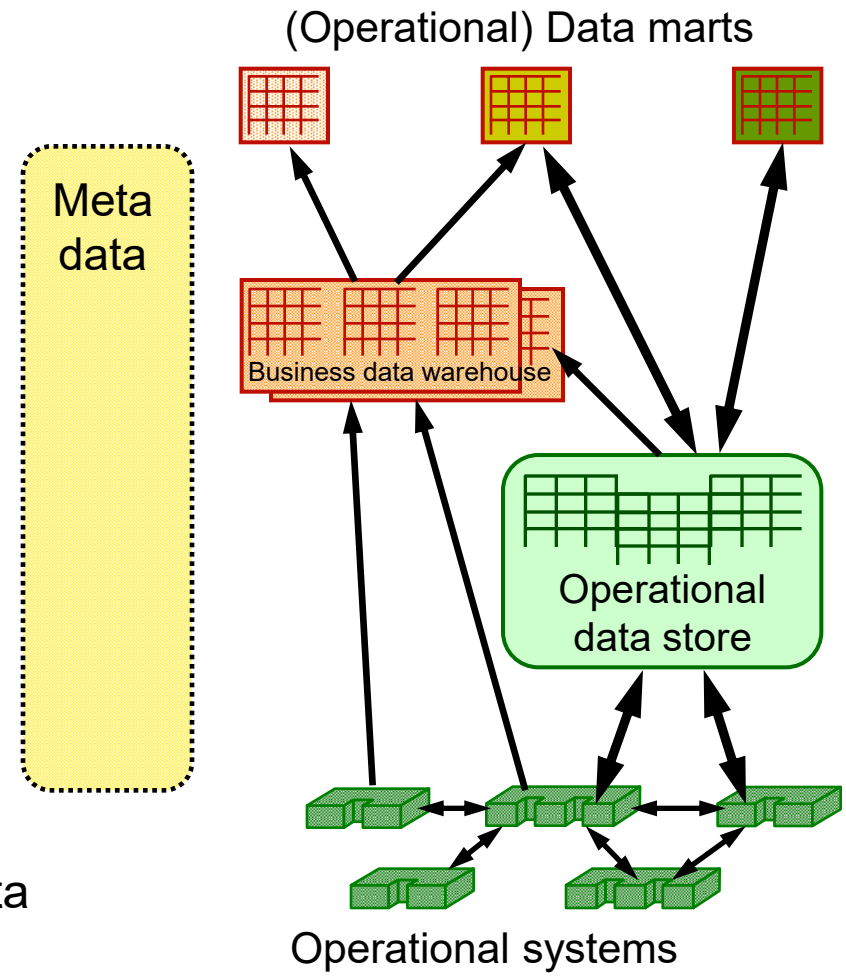
- Business data warehouse
  - Reconciling disparate data
  - **Single version of the truth**
  - Creates historical record
- Characteristics
  - Historical data
  - Separation of informational & operational needs
  - Structured data
  - Unidirectional data flow
  - Trusted sources
  - Basic technical metadata

A man with one watch knows what time it is.
A man with two watches is never sure.
Segal's Law



Late-1990s: Operational data stores and data marts

- Operational data store (ODS)
  - Near real-time
- Integrating related data
  - Relative/partial truth
- Characteristics
  - Recent and historical data
  - Merging of informational & operational needs
  - Structured data (largely)
  - Bidirectional data flow
  - Trusted sources
  - Comprehensive technical metadata



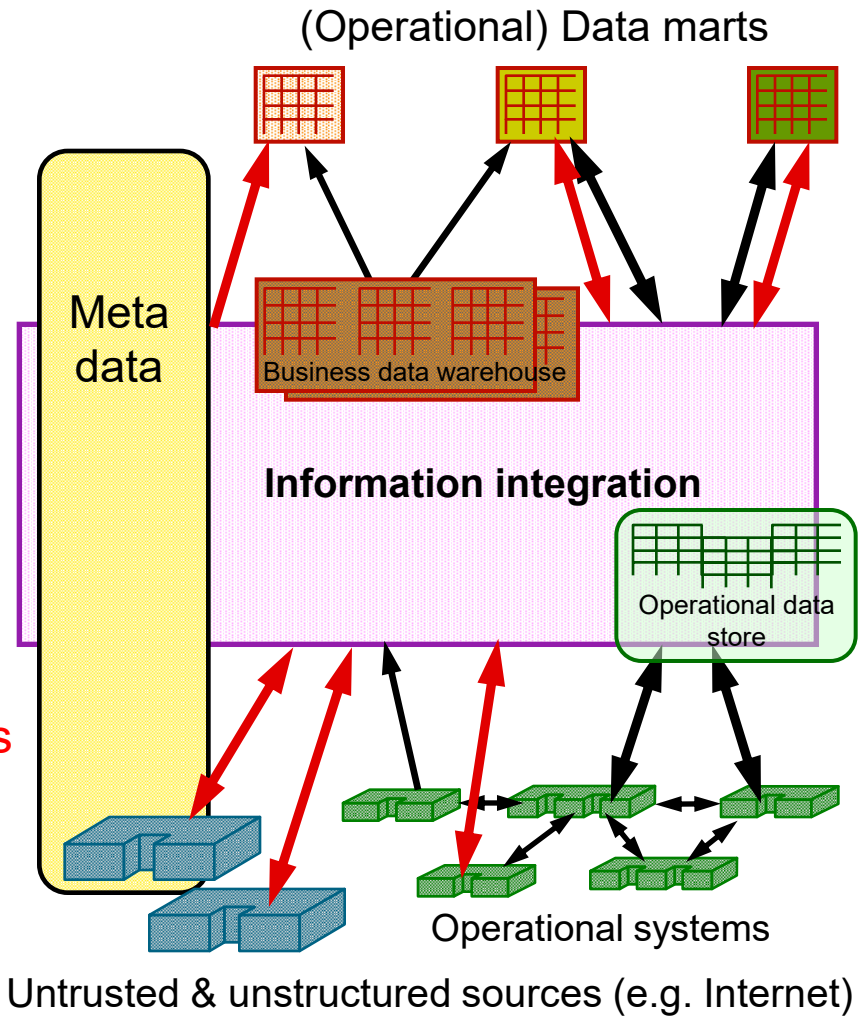
Today and future business needs encompass a fully integrated approach to management and operation.

- Support for decision making at every level
  - Strategic and tactical, sometimes automated, real-time
  - Individual and group decision making
- Comprehensive business information
  - Historical, point-in-time and up-to-the-second information
  - Plus extensive external data, unstructured data etc.
- Aligned across different departments
  - And geographies, companies, competitors, customers, markets
- Integrated into the overall business process
  - True linkage to real-time operational systems
  - Within the organization and with partners, customers, etc.

Today – this is real business intelligence!

Today: Comprehensive integration of information

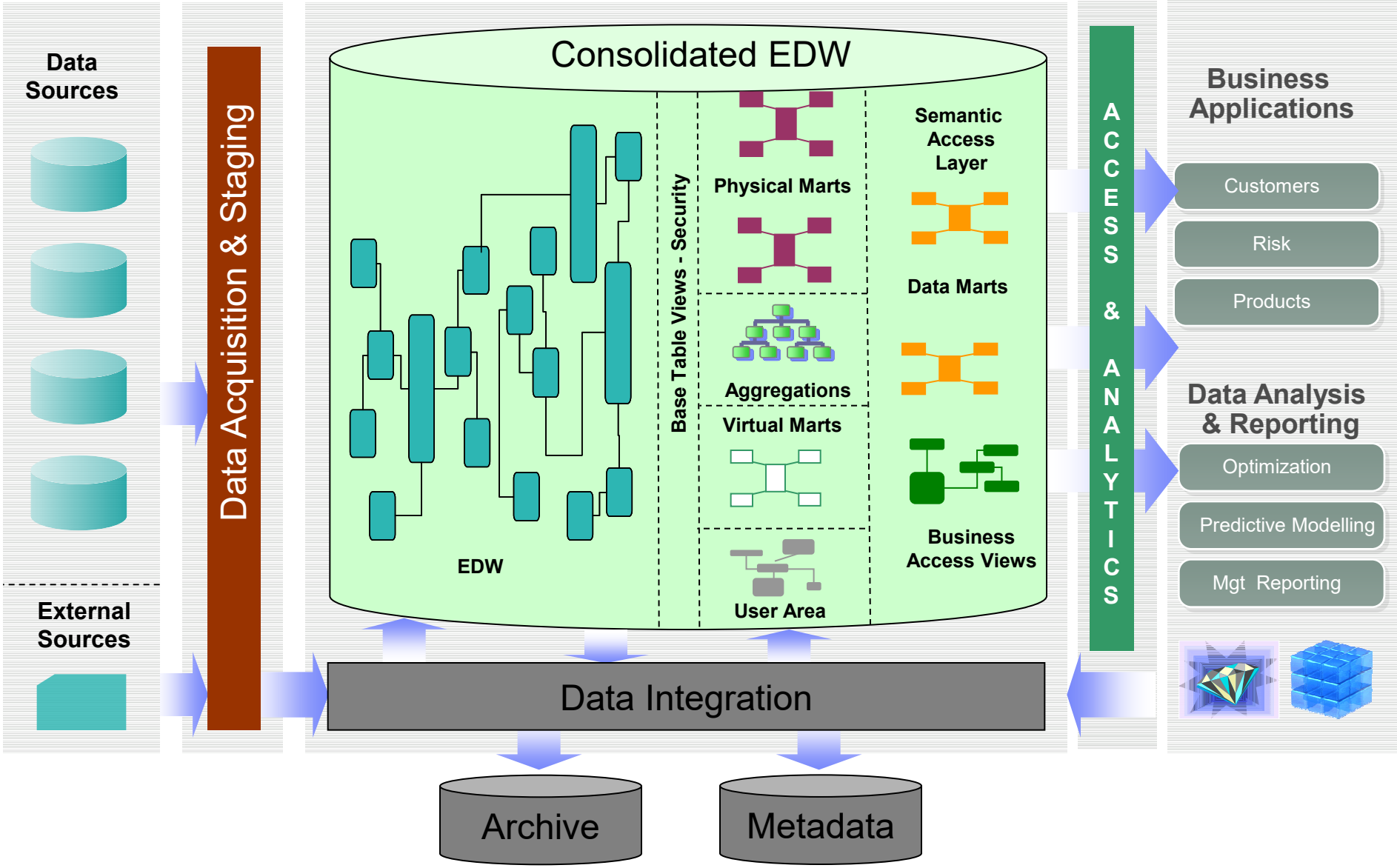
- Integrated information
  - Real-time knowledge
  - Integrating all information
  - Complete truth
- Characteristics
  - Immediate and historical data
  - Fully merged informational & operational needs
  - Structured and unstructured data
  - Bidirectional data flow and access
  - Intelligent caching
  - Trusted and untrusted sources
  - Complete business & technical metadata



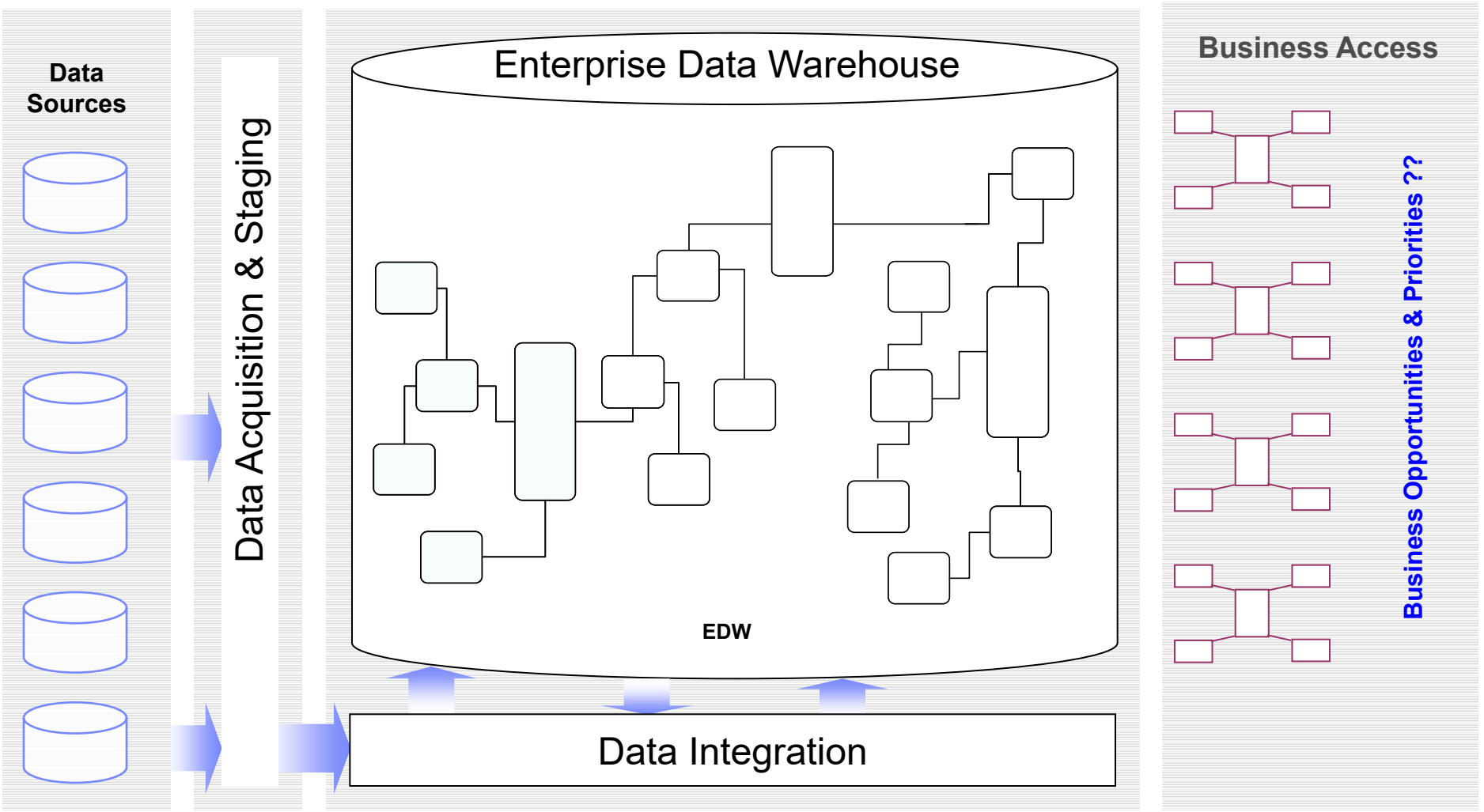
Data Warehouse Architecture

ATVANTAGE
TIMETOACT GROUP

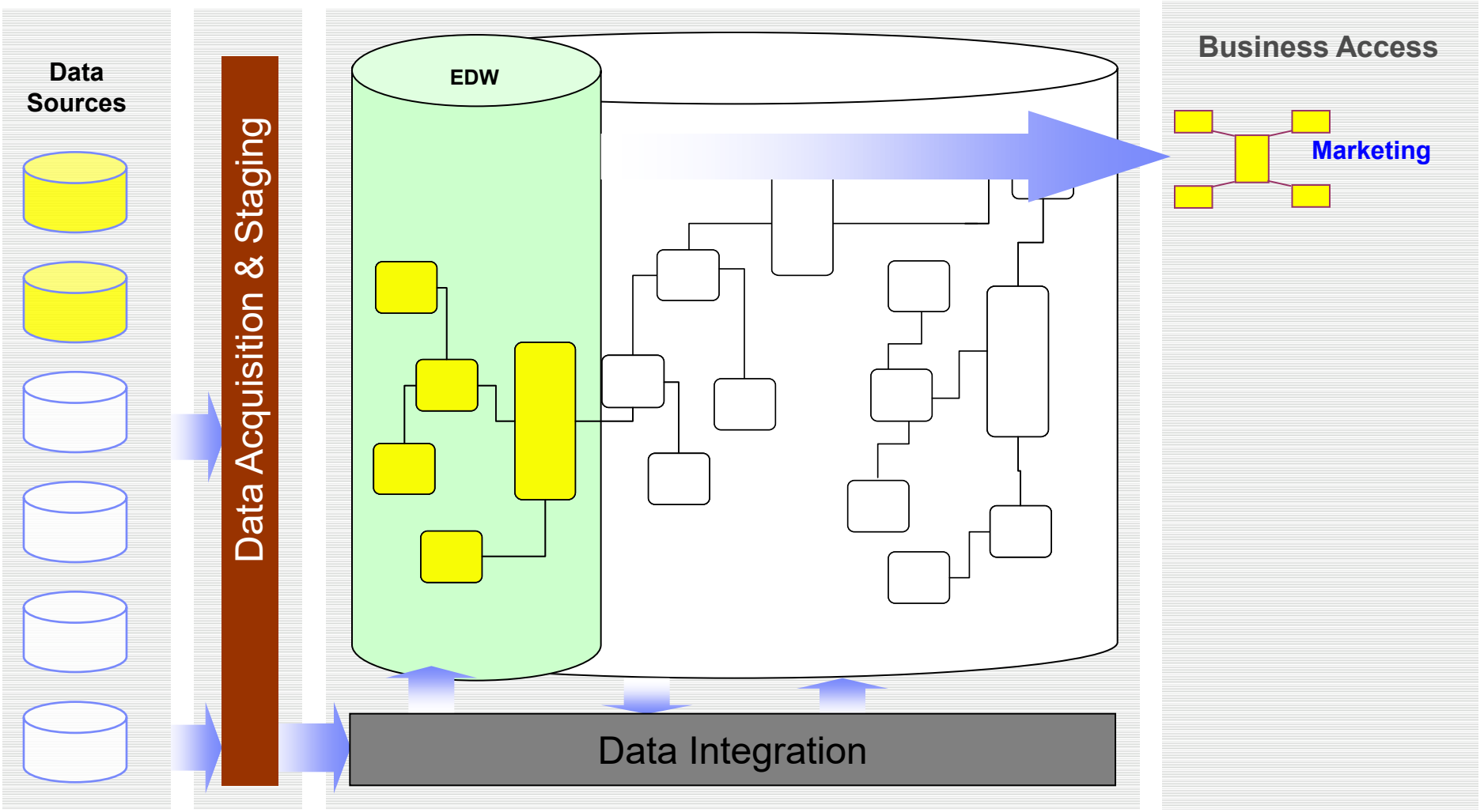
Layered Data Architecture for Data Warehousing



Enterprise Data Warehouse – Building Block Process

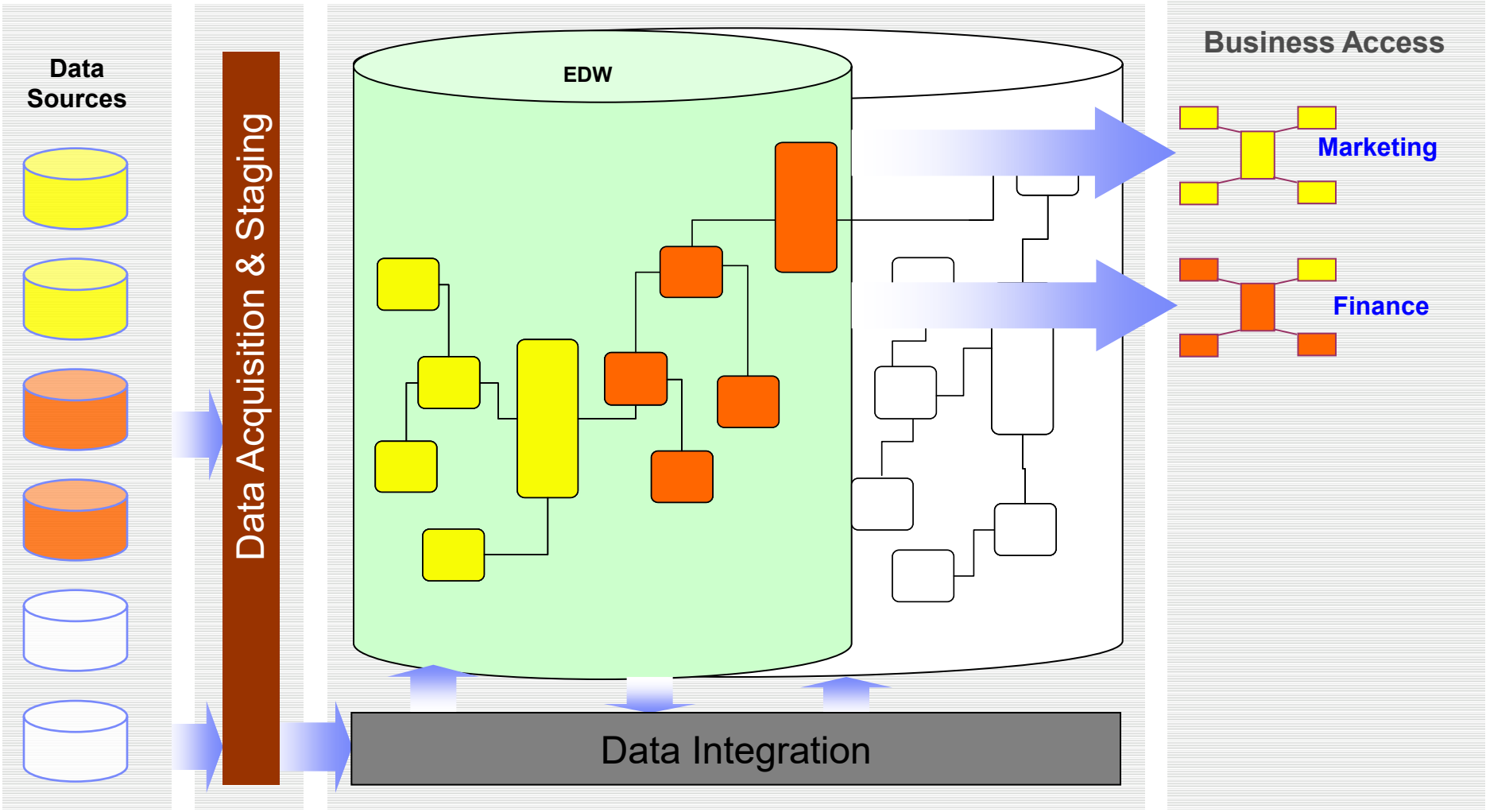


Enterprise Data Warehouse – Building Block Process



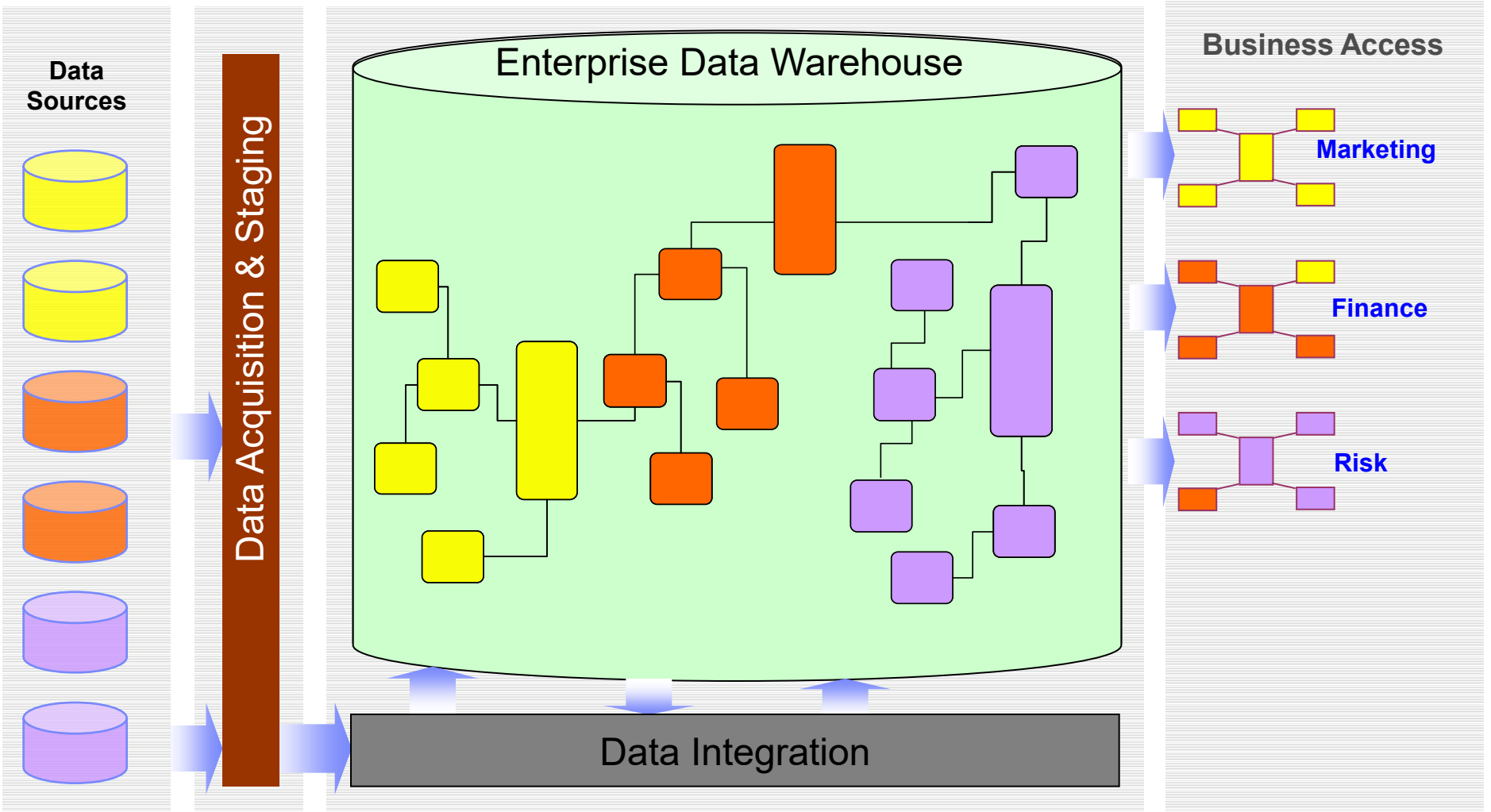
Leverage Data Loads – Load Once and Use by Many

Enterprise Data Warehouse – Building Block Process



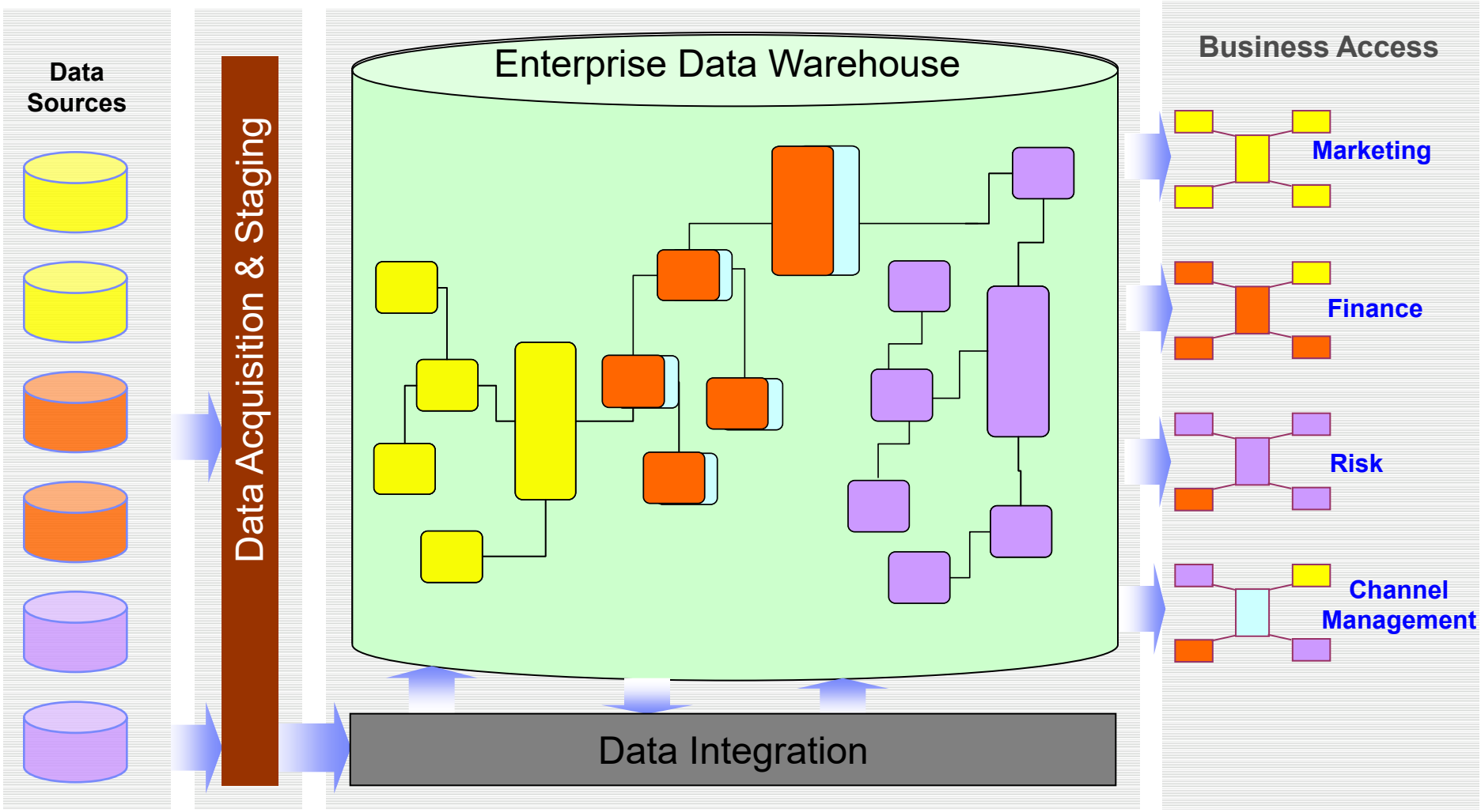
Leverage Data Loads – Load Once and Use by Many

Enterprise Data Warehouse – Building Block Process



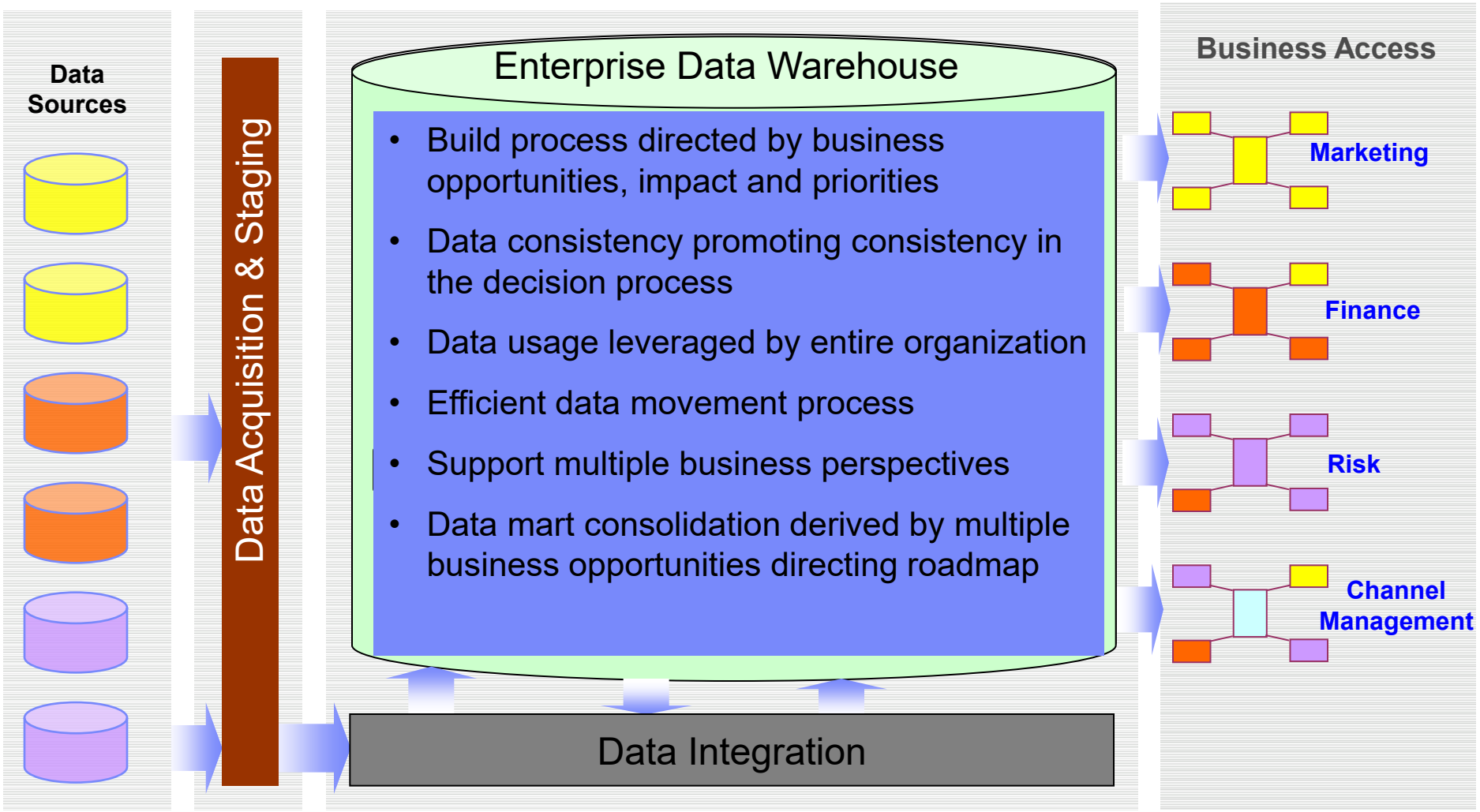
Leverage Data Loads – Load Once and Use by Many

Enterprise Data Warehouse – Building Block Process



Leverage Data Loads – Load Once and Use by Many

Enterprise Data Warehouse – Building Block Process



Leverage Data Loads – Load Once and Use by Many

Dimensional Data Modeling

ATVANTAGE
TIMETOACT GROUP

Data Modeling for OLTP Systems

- Requirements
  - Efficient update operations
  - Efficient read operations
  - As little redundancy as possible
  - Easy maintenance of the data model

→As little redundancy as possible in the data model

Codd‘s normal forms for database relations (1)

- First Normal Form (1NF):
  - Every table has a minimal set of key attributes that can uniquely identify a record.
  - No field values can be sets, i.e., only single values are allowed

CD_ID	Album	Jahr der Gründung	Titelliste
4711	Anastacia - Not That Kind	1999	{1. Not That Kind, 2. I'm Outta Love, 3. Cowboys & Kisses}
4712	Pink Floyd - Wish You Were Here	1964	{1. Shine On You Crazy Diamond}
4713	Anastacia - Freak of Nature	1999	{1. Paid my Dues}

(Example from Wikipedia)

CD_ID	Albumtitel	Interpret	Jahr der Gründung	*Track*	Titel
4711	Not That Kind	Anastacia	1999	1	Not That Kind
4711	Not That Kind	Anastacia	1999	2	I'm Outta Love
4711	Not That Kind	Anastacia	1999	3	Cowboys & Kisses
4712	Wish You Were Here	Pink Floyd	1964	1	Shine On You Crazy Diamond
4713	Freak of Nature	Anastacia	1999	1	Paid my Dues

Codd‘s normal forms for database relations (2)

- Second Normal Form (2NF):
  - In 1st normal form
  - Every non-key attribute is fully dependent on the key.
    There are no dependencies between a partial key and a non-key field.

CD_ID	Albumtitel	Interpret	Jahr der Gründung	*Track*	Titel
4711	Not That Kind	Anastacia	1999	1	Not That Kind
4711	Not That Kind	Anastacia	1999	2	I'm Outta Love
4711	Not That Kind	Anastacia	1999	3	Cowboys & Kisses
4712	Wish You Were Here	Pink Floyd	1964	1	Shine On You Crazy Diamond
4713	Freak of Nature	Anastacia	1999	1	Paid my Dues

CD_ID	*Track*	Titel
4711	1	Not That Kind
4711	2	I'm Outta Love
4711	3	Cowboys & Kisses
4712	1	Shine On You Crazy Diamond
4713	1	Paid my Dues

CD_ID	Albumtitel	Interpret	Jahr der Gründung
4711	Not That Kind	Anastacia	1999
4712	Wish You Were Here	Pink Floyd	1964
4713	Freak of Nature	Anastacia	1999

Codd‘s normal forms for database relations (3)

- Third Normal Form (3FN):
  - In 2nd normal form
  - No functional dependencies between non key fields.

CD_ID	Albumtitel	Interpret	Jahr der Gründung
4711	Not That Kind	Anastacia	1999
4712	Wish You Were Here	Pink Floyd	1964
4713	Freak of Nature	Anastacia	1999

CD_ID	Albumtitel	Interpret
4711	Not That Kind	Anastacia
4713	Freak of Nature	Anastacia
4712	Wish You Were Here	Pink Floyd

Interpret	Jahr der Gründung
Anastacia	1999
Pink Floyd	1964

Charateristics of OLTP Data Models

- Lots of joins necessary to answer complex questions
- Lots of tables
- Lots of associations between tables
- Complex structure, not easy to understand

Data Modeling for a Data Warehouse

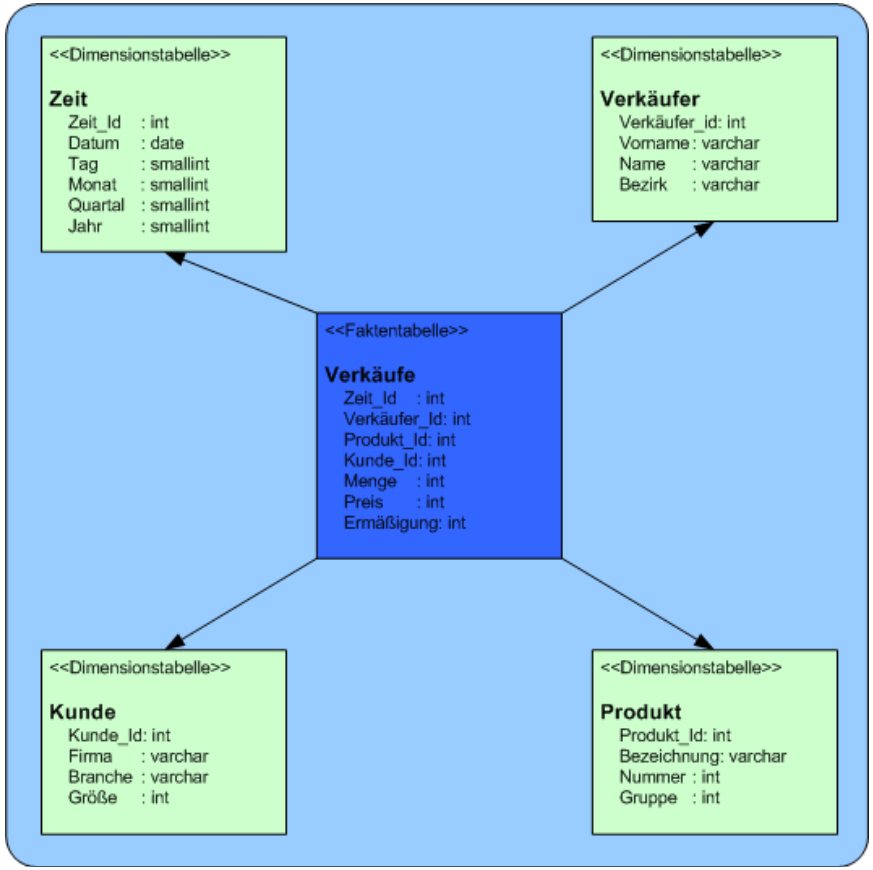
- Multi-dimensional data model
  - Key Figures (i.e. sales amount, profit) – “measures”
  - Dimensions
    - Attributes like
      - Product
      - Region
      - Time period (day, week, month, year)
    - Correspond to subjects
  - For every combination of dimension attribute values one value of each key figure
    - Sales amount for product X in region y and time period z

Multidimensional data model

- Dimension tables
  - Table for each dimension like product, region, time period
  - Primary key identifies each dimension element
  - Additional fields contain descriptive information like product name
- Fact tables
  - Primary key of dimension tables are foreign keys
  - The other fields contain the values of the key figures/measures

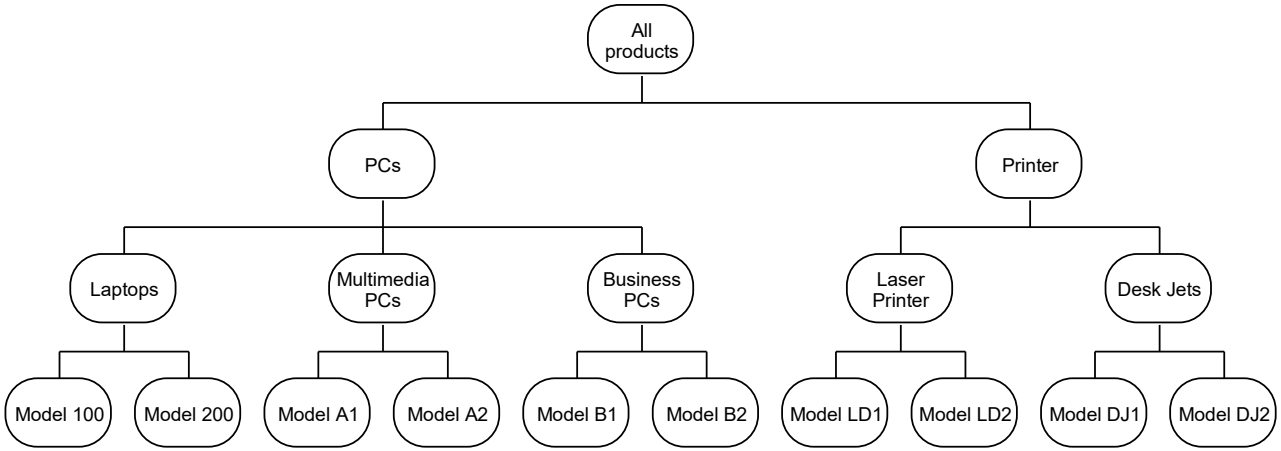
Star Schema

- Example from German Wikipedia



Hierarchies

- Dimensions can be organized as hierarchies
  - i.e. product hierarchy



Hierarchies

- Other hierarchies:
  - Date → Month/Year → Quarter/Year →Year
  - Customer → Company → Industry
  - City → County → State → Country → Continent
- Arbitrary number of hierarchy levels
- Purpose:
  - group and structure data
  - enable view on data on differently detailed levels
- Hierarchies define aggregations
  - aggregated values for measures for each hierarchy level

Data Models for Hierarchies

- Denormalized
  - 1 Table with all hierarchy levels
  - Advantage
    - Efficent aggregations
  - Disadvantage
    - Complex updates if hierarchies change

Productid	Productname	Productgroup	Productcategory	Productclass
1234ABC	Thinkpad T60	Laptop	PC	Computer

Data Models for Hierarchies

- Normalized
  - 1 table for each hierarchy level
- Advantage
  - Minimal updates for changes in the hierarchies
- Disadvantage
  - More complex queries when computing aggregations
    - Multiple joins

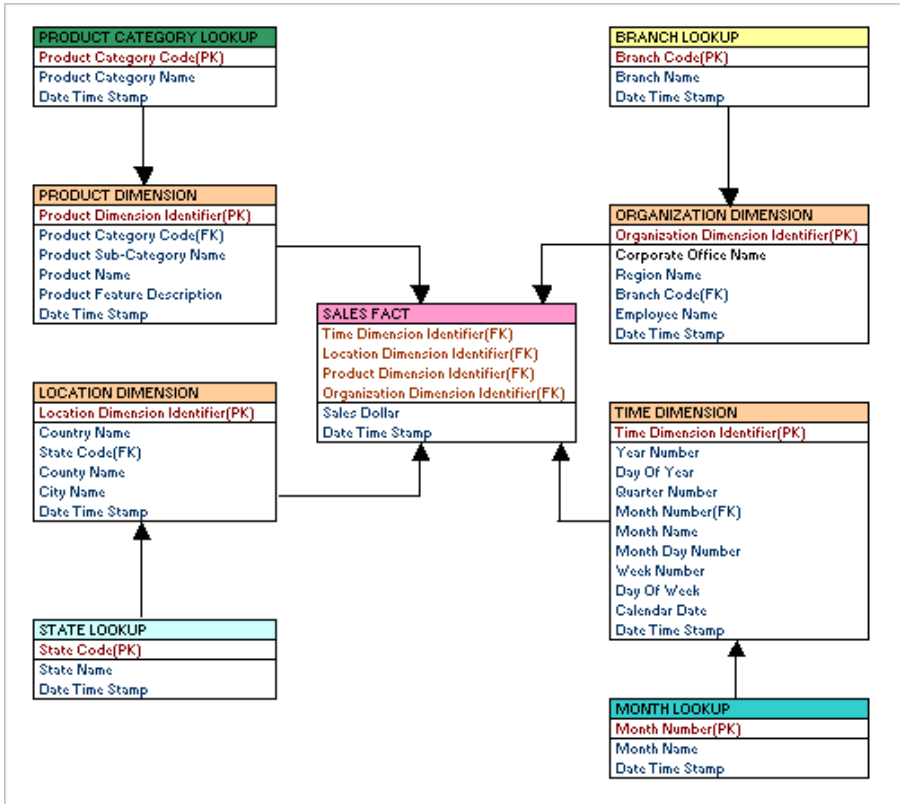
Productid	Productname	Productgid
1234ABC	ThinkPad T60	G1234

Productgid	Productgroup	Productcatid
G1234	Laptop	CAT12

Productclassid	Productclass
C3	Computer

Snowflake Schema

- Multiple dimension tables for hierarchies
  - mostly normalized



Multidimensional Operations

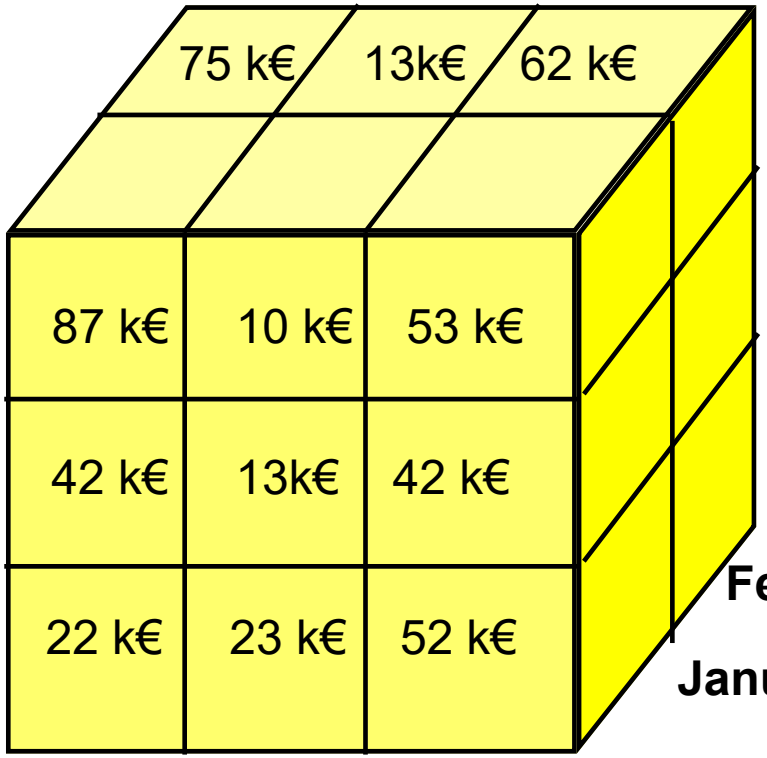
ATVANTAGE
TIMETOACT GROUP

Multidimensional data model

- Dimension tables
- Fact tables

→ This constitutes a virtual „data cube“.

Example:



City: Berlin

Munich

Stuttgart

Year/Month:

February 2006

January 2006

Product:

TP T60 TP R50 TP Z61

Multidimensional data model

- Advantages
  - Lots of information provided "at a glance"
  - Information of one type grouped
  - Performance optimized for read-operations on multi-dimensional array
    - DB optimizations exist for such data models
      - Multi dimensional clustering
      - Column-based tables or indexes
      - Star Join Indexes
      - Etc.

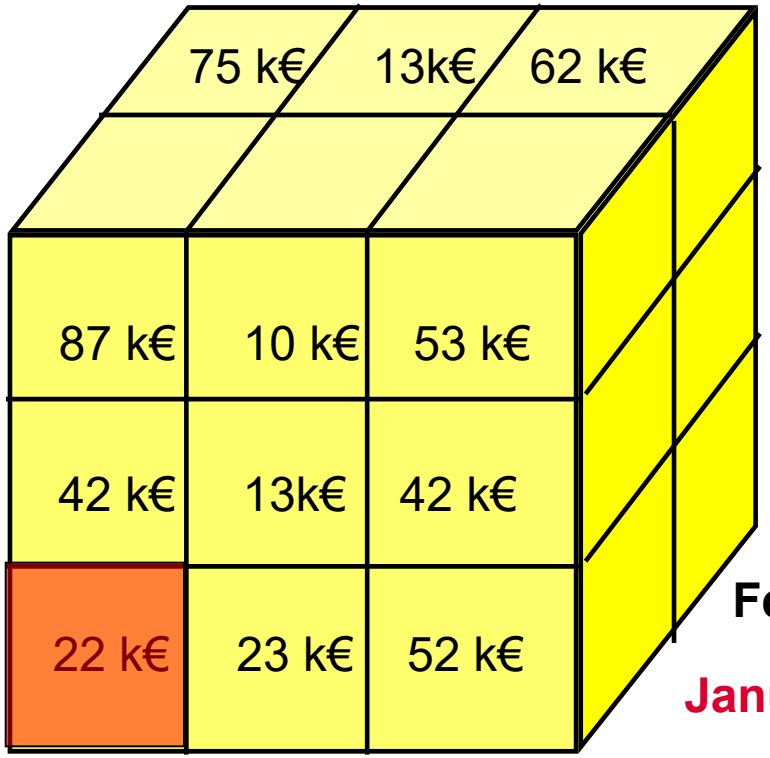
Multdimensional data model

- Disadvantages of a multidimensional table
  - High Redundancy
  - Sparsity
    - Space reserved for each possible value (each combination of dimensions)
    - "null" is stored in a field with same length as any value
  - Limited size
    - more dimensions result in much higher data volume
    - exponential growth

Multi-dimensional operations

- Selection
  - Definition of a filter
  - Select data of a single cell with a condition for each dimension
    - For instance:
      - time = 'January 2006'
      - location = 'Stuttgart'
      - product = ‘ThinkPad T60'

Example:



City: Berlin

Munich

Stuttgart

Year/Month:

February 2006

January 2006

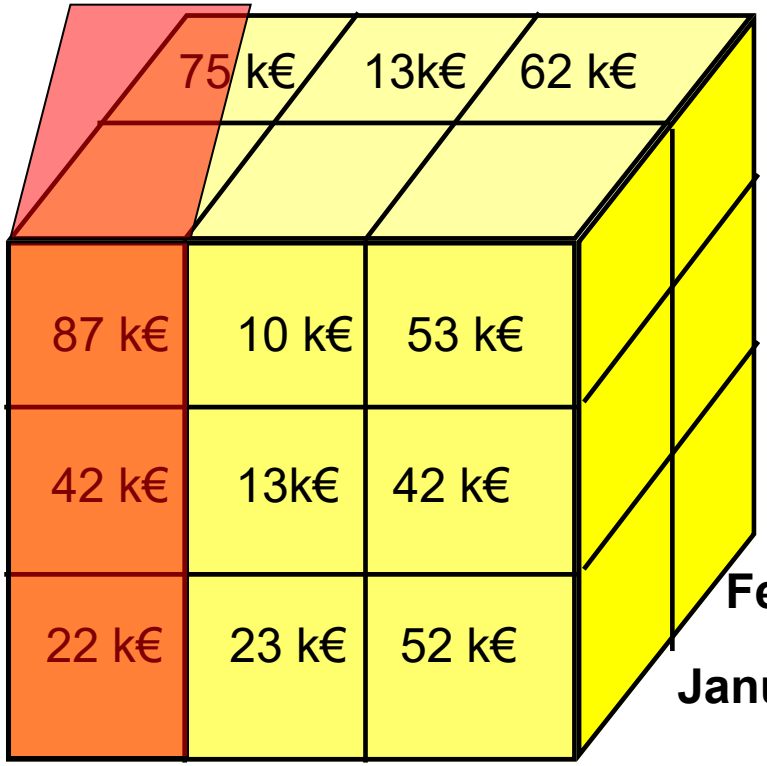
Product:

TP T60 **TP R50** **TP Z61**

Multi-dimensional operations

- Slice
  - Definition of a filter
  - Select a "slice" from the "data cube"
  - Condition for one single dimension
  - For instance
    - Product = ‘ThinkPad T60'

Example:



City: Berlin

Munich

Stuttgart

Year/Month:

February 2006

January 2006

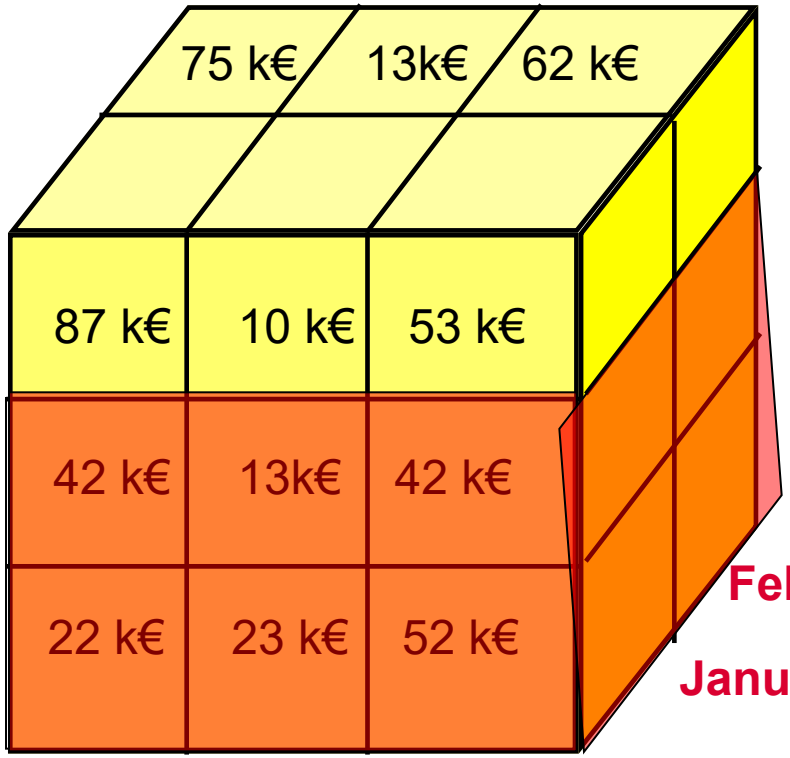
Product:

TP T60 **TP R50** **TP Z61**

Multi-dimensional operations

- Dice
  - Definition of intervals/sets as filter
  - Select a smaller cube
  - Conditions for instance
    - time = 1st quarter (January, February, March)
    - location = region south (Stuttgart, Frankfurt, Munich)

Example:



City: Berlin

Munich

Stuttgart

Year/Month:

February 2006

January 2006

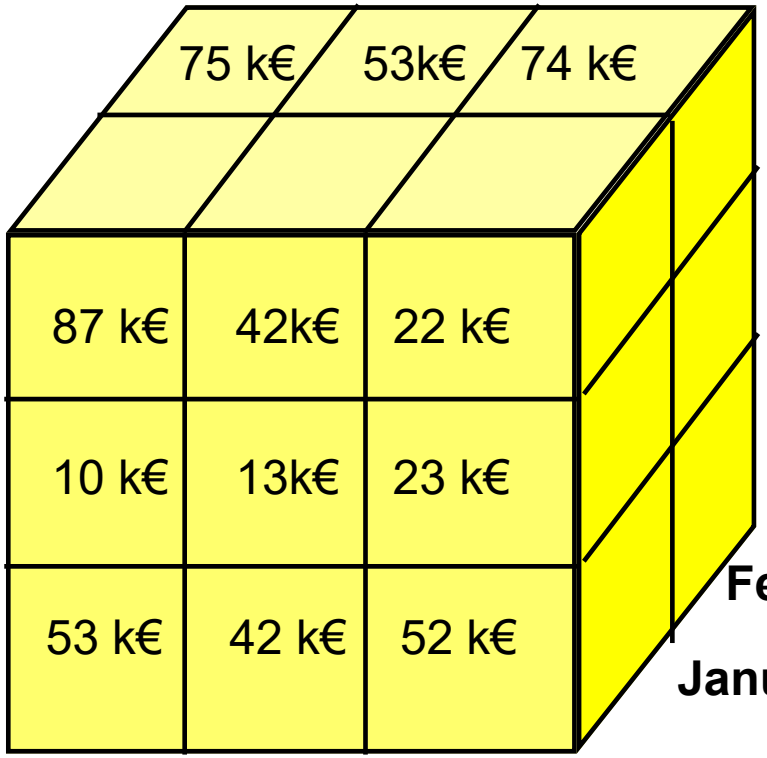
Product:

TP T60 TP R50 TP Z61

Multi-dimensional operations

- Rotate/Pivot
  - Rotate cube along its axes
  - Get different view on data cube
  - # of views on cube = (# of dimensions)!
    - 2 dimensions, 2 views
    - 3 dimensions, 6 views
    - 4 dimensions, 24 views
    - ...

Example:



City→
Product:

TP T60

TP R50

TP Z61

Year/Month:

February 2006

January 2006

Product → City:

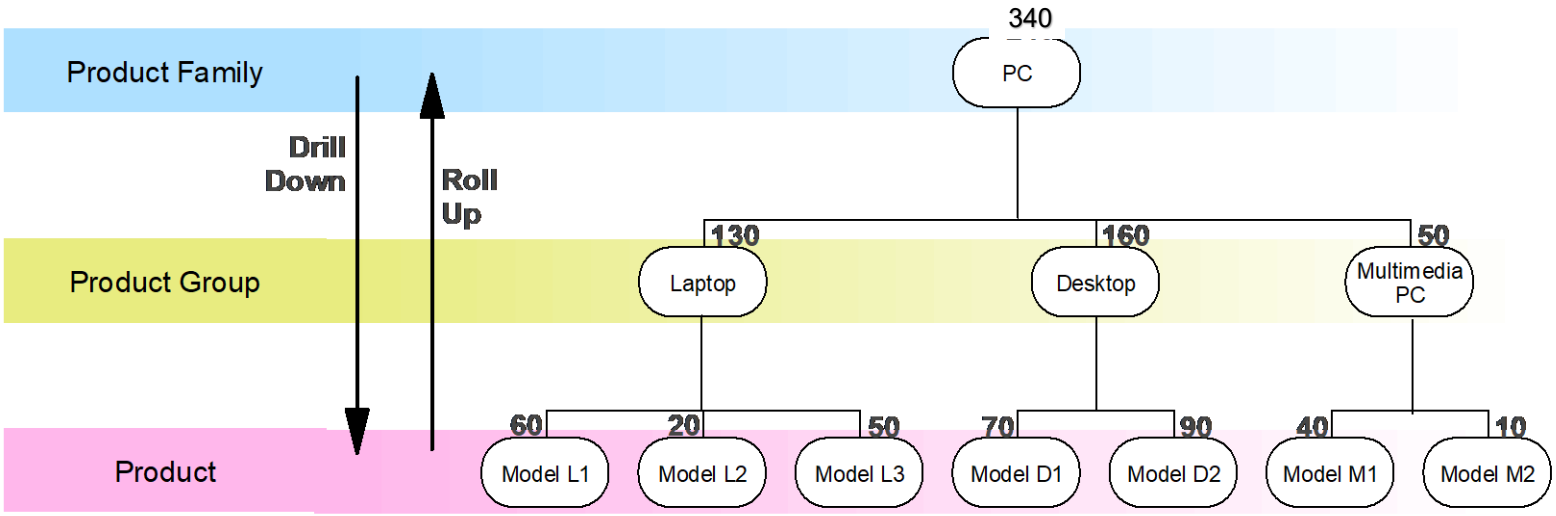
Berlin Munich Stuttgart

Multi-dimensional operations

- Roll-up & Drill-down
  - Prerequisites:
    - Hierarchies defined
    - Aggregated data for all hierarchy levels available
  - Roll up: change hierarchy level "upwards":
    - get less detailed data
  - Drill down: change hierarchy level "downwards":
    - get more detailed data

Multi-dimensional operations

- Roll-up & Drill-down



Presentation of multi-dimensional data

- Pivot Tables

		January	February	March	1st Quarter Sum
Computer	Stuttgart	11	10	12	33
	Frankfurt	8	9	14	31
	Munich	10	9	10	29
	Sum	29	28	36	93
Accessories	Stuttgart	9	9	11	29
	Frankfurt	5	6	4	15
	Munich	7	8	9	24
	Sum	21	23	24	68
Sum		50	51	60	161

Exercise

- For one of the following companies
  - Retail Bank
  - Telecommunication company
  - Online bookstore (like Amazon.com)
  - Discount furniture store (like IKEA)
  - Supermarket

1. Outline the data model for a data warehouse for a specific subject area of your choice
2. For 2-3 sample questions outline the processing of the queries (for instance which tables need to be joined or which aggregations have to be computed).