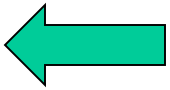
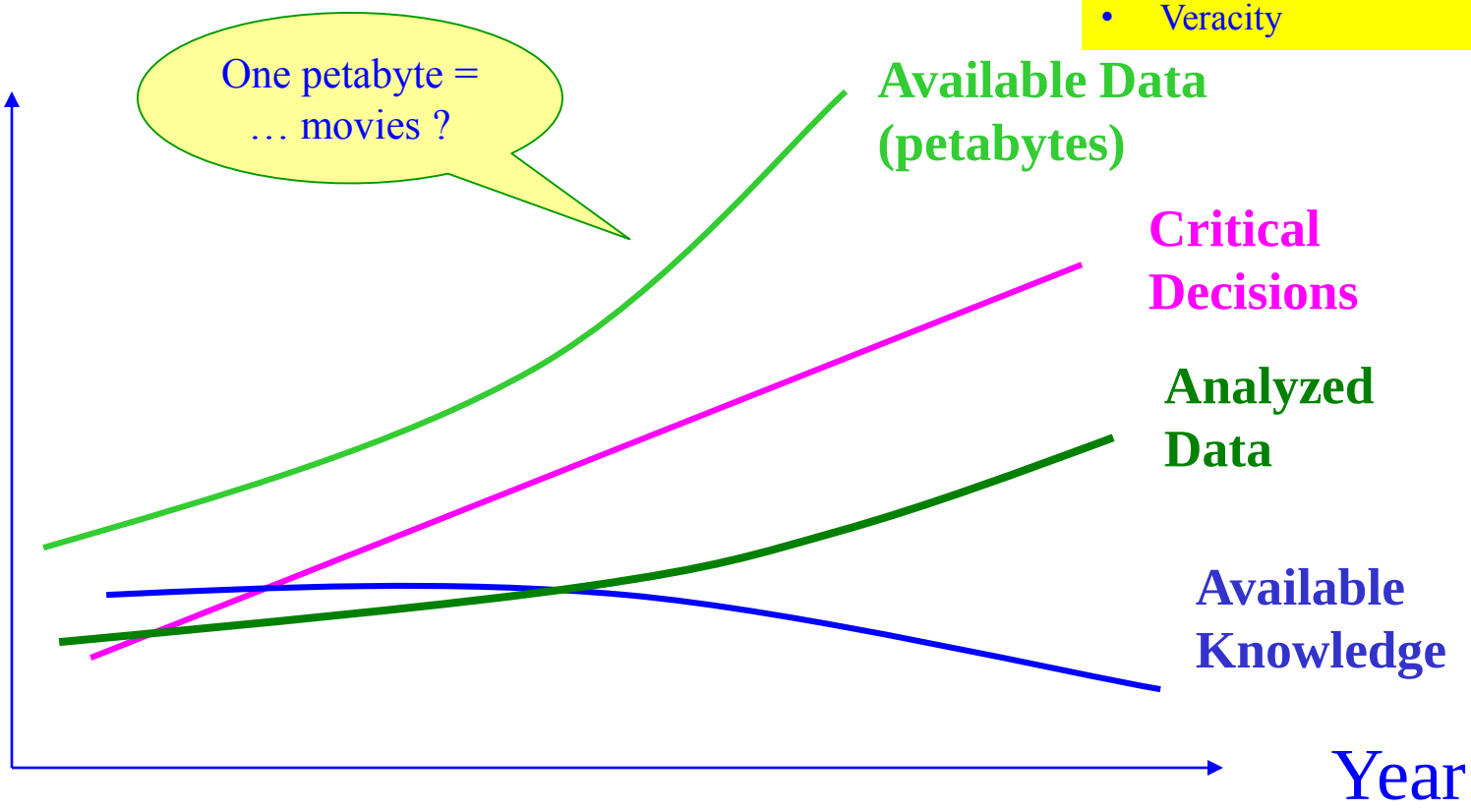
Lecture No. 12 – Business Intelligence and Data Warehousing

- Business Intelligence 
- Why Data Warehousing?
- Data Modeling
- Metadata ("data about data")
- Data Quality
- ETL (Extraction, Transformation, and Loading)

The Facts Gap

The FOUR V's of Big Data
- Volume
- Variety
- Velocity
- Veracity



Example: Customer Relationship Management (CRM)

From: http://megaslides.com/doc/892947/facts-about-customer-relations

- Cost of selling to a new customer is six times as high as to existing customer
- Odds of selling to a new customer = 1/7 to an existing customer = 1/2
- Each dissatisfied customer tells 8 to 10 people
- 70% of dissatisfied customers will do business again if they feel their complains are handled well
- 1 extra % of customer retention can boost turnover by as much as 15%

Business Intelligence (BI)

- "The processes, technologies and tools needed to turn data into information and information into knowledge and knowledge into plans that drive profitable business action. BI encompasses data warehousing, business analytics and knowledge management.

The Data Warehouse Institute, Q4/2002

What is Business Intelligence?

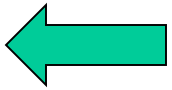
- Relationship of intelligence to various levels of summarisation
  - Data – unstructured data
  - Information – structured data useful for analysis
  - Knowledge - obtained from experts based on actual experience
  - Intelligence – keen insight into understanding important relationships

Thierauf (2001)

Role of BIS

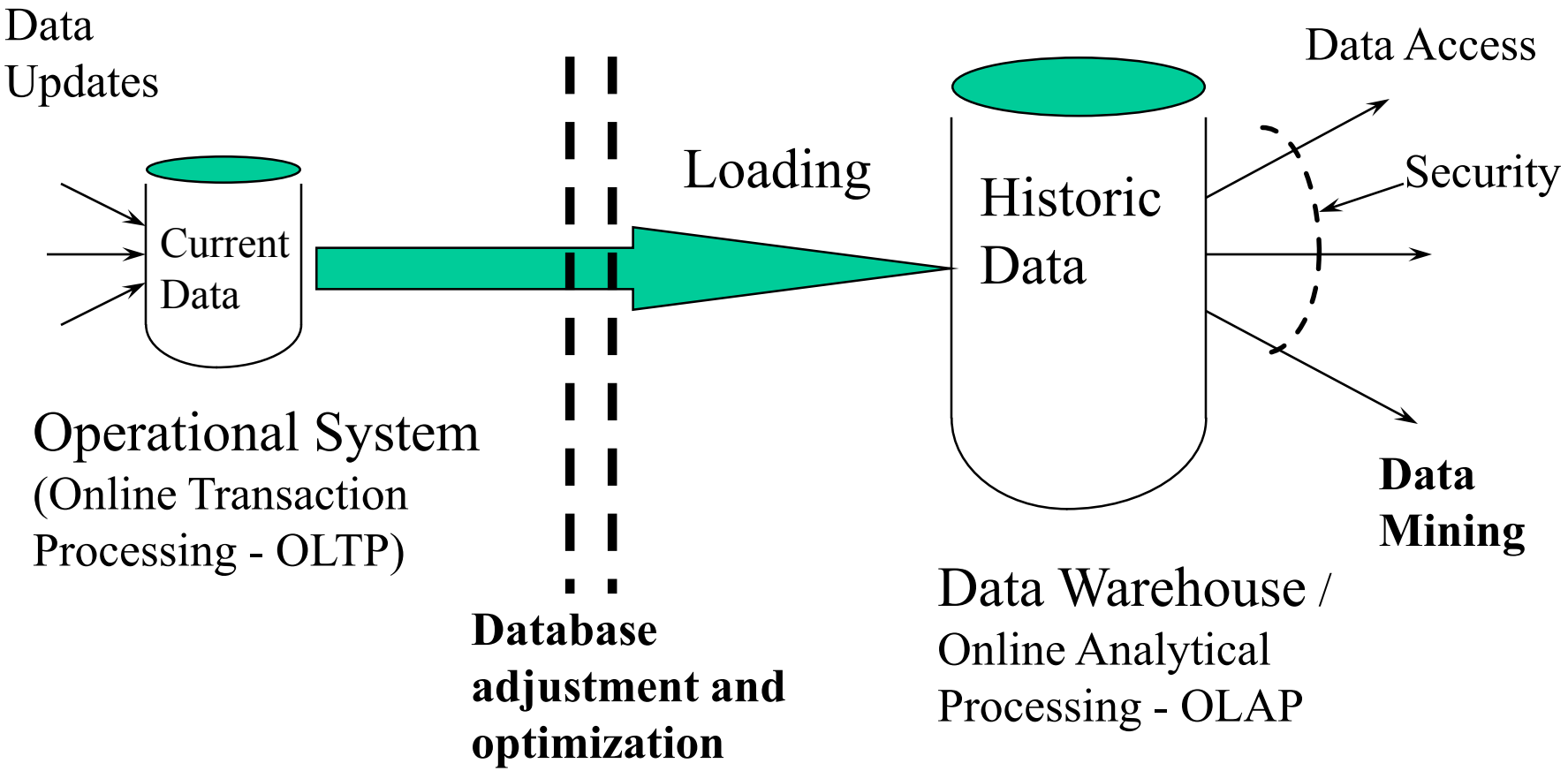
- Provide decision makers with timely data, information and knowledge for problem solving, and problem finding
- Past : Decision making as Problem Solving activity
  - *Reactive approach –use of appropriate management technologies to resolve current problems as they arise*
- Current: Business intelligence activity as problem solving, as well as problem finding
  - *Proactive, preventive approach – anticipating future company problems; looking for future opportunities*

Lecture No. 12 – Business Intelligence and Data Warehousing

- Business Intelligence
- Why Data Warehousing? 
- Data Modeling
- Metadata ("data about data")
- Data Quality
- ETL (Extraction, Transformation, and Loading)

Data Warehouse Concept

(Example: CRM)



Data Warehouse - Definitions

- A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making - Inmon, 1994
- A read-only analytical database that is used as the foundation of a decision support process - Poe and Reeves, 1995
- Managed data situated after and outside the operational systems - Gupta, 1997
- A few terabytes of data stored in a dark and cool place - Last, 1998

DW - Characteristics

- **Subject-Oriented**: production floor control, marketing, purchasing, QC, customer support, etc.
- **Integrated**: encoding, measurement units, naming conventions, key structures.
- **Time-Variant**: long time horizon, time keys.
- **Nonvolatile**: no updates of data (only data loading and data access).

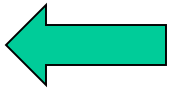
OLTP vs. DWH

	On-Line Transaction Processing	Data Warehouse
Users	Front-line workers	Management
Purpose	Supports day-to-day operations	Supports strategic decisions
Data	Raw data (entered by users)	Filtered and transformed data
Source of data	Internal sources only	Internal and external sources
Time horizon	Current data	Historical data
Level of detail	Only detail data	Detail and summary data
Data structure	3NF (Why?)	De-normalized tables
Design goal	Maximum update efficiency	Maximum query efficiency

High concurrency
Low latency

Low concurrency
Relaxed latency

Business Intelligence and Data Warehousing

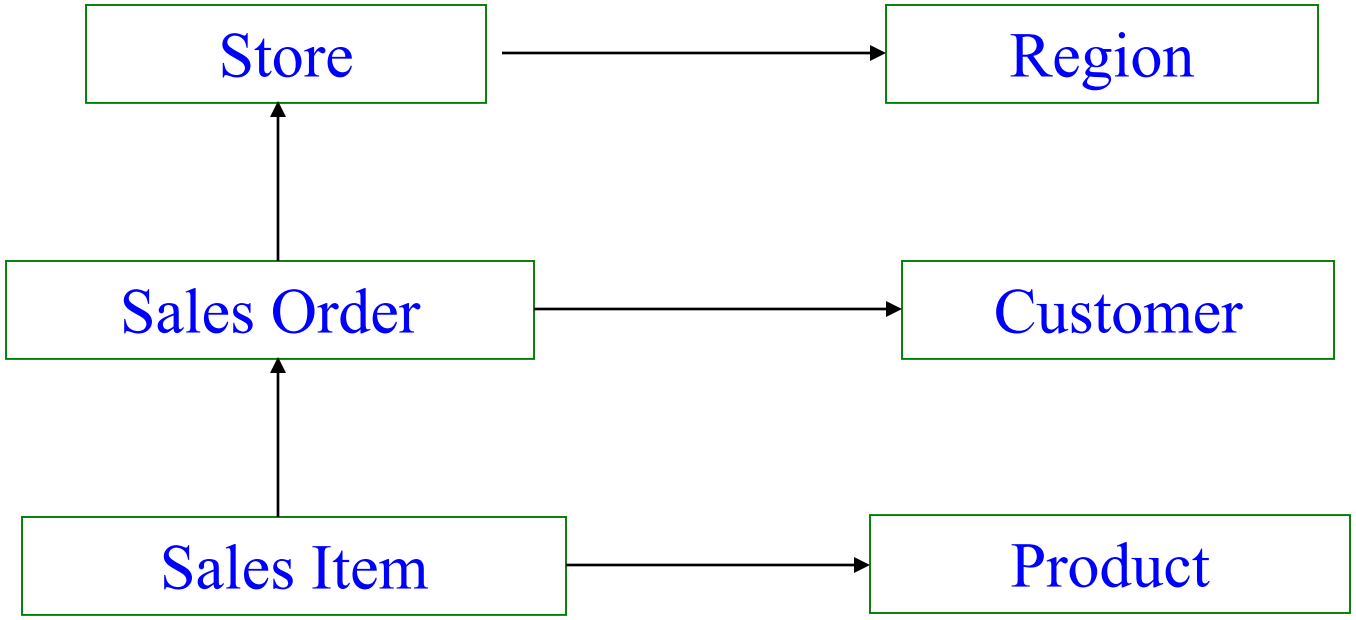
- Business Intelligence
- Why Data Warehousing?
- Data Modeling 
- Metadata ("data about data")
- Data Quality
- ETL (Extraction, Transformation, and Loading)

Data Models

- What is a data model?
  - The perception of the data (at different levels of abstraction) by users, designers, and developers of an information system
  - The same data may be modeled in many different ways!
- The ER / Relational Model
  - The business keeps track of data about *entities*
  - Entities are stored in *tables*
  - All entities are created equal
- The Dimensional (Star-Join) Model
  - A few entities (called “facts”) are much more important than the other entities

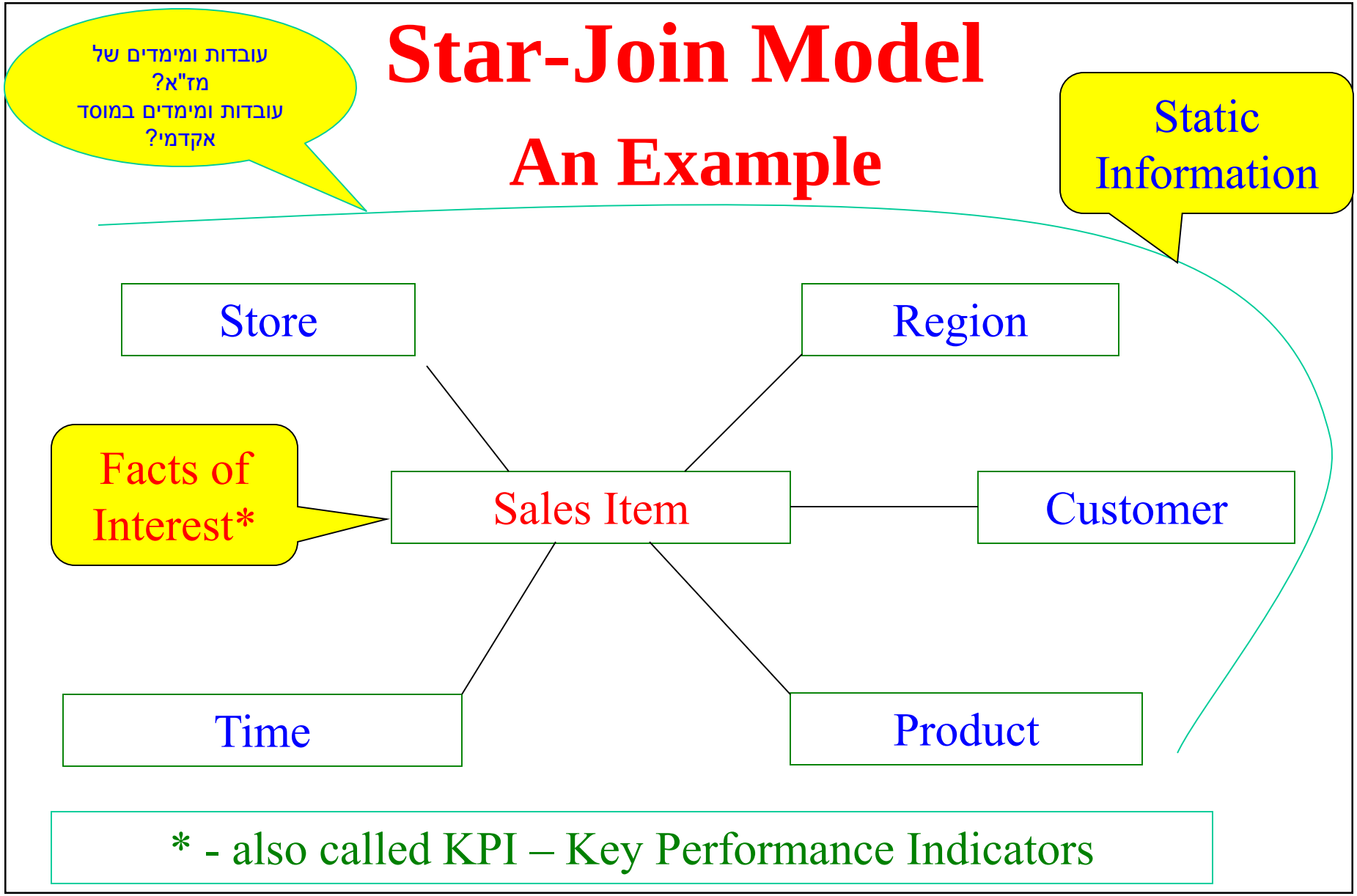
Entity-Relationship Model

An Example



Star-Join Model

An Example



Multidimensional Logical Model / Star Schema Analysis

Definition 1

A **dimension** is a logical grouping of attributes arranged according to business area

- **Examples**: Customer, Product, Location, and Time
- Most dimensions are represented by *descriptive* values (e.g., customer name, product code, order number, etc.)
- Other names for **descriptive**: qualitative, categorical, nominal, non-ordinal.
- Dimensional data (e.g., information about customers) is stored in *reference entities*

Multidimensional Logical Model / Star Schema Analysis (cont.)

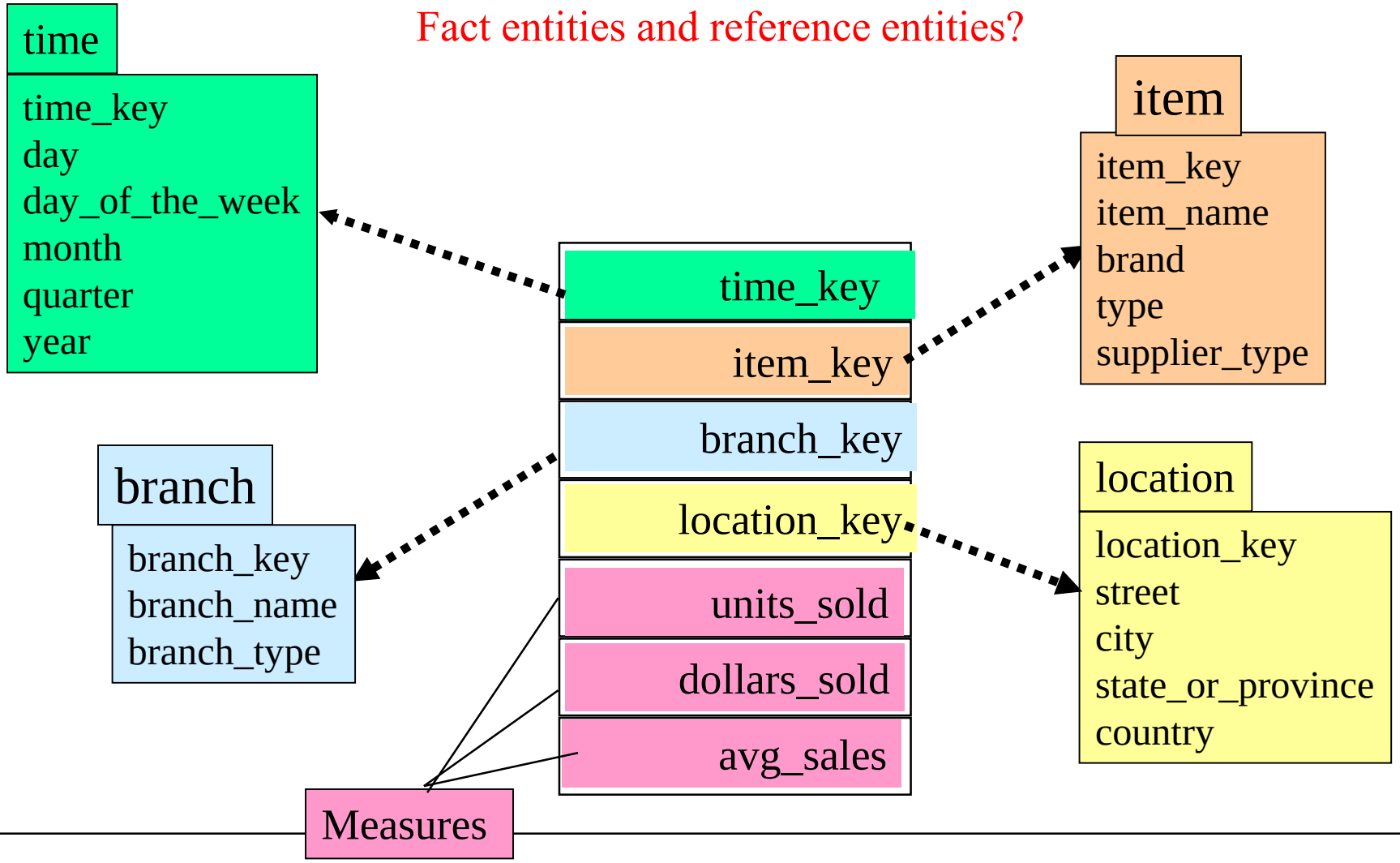
Definition 2

Facts (business metrics) - points of dimensional intersection

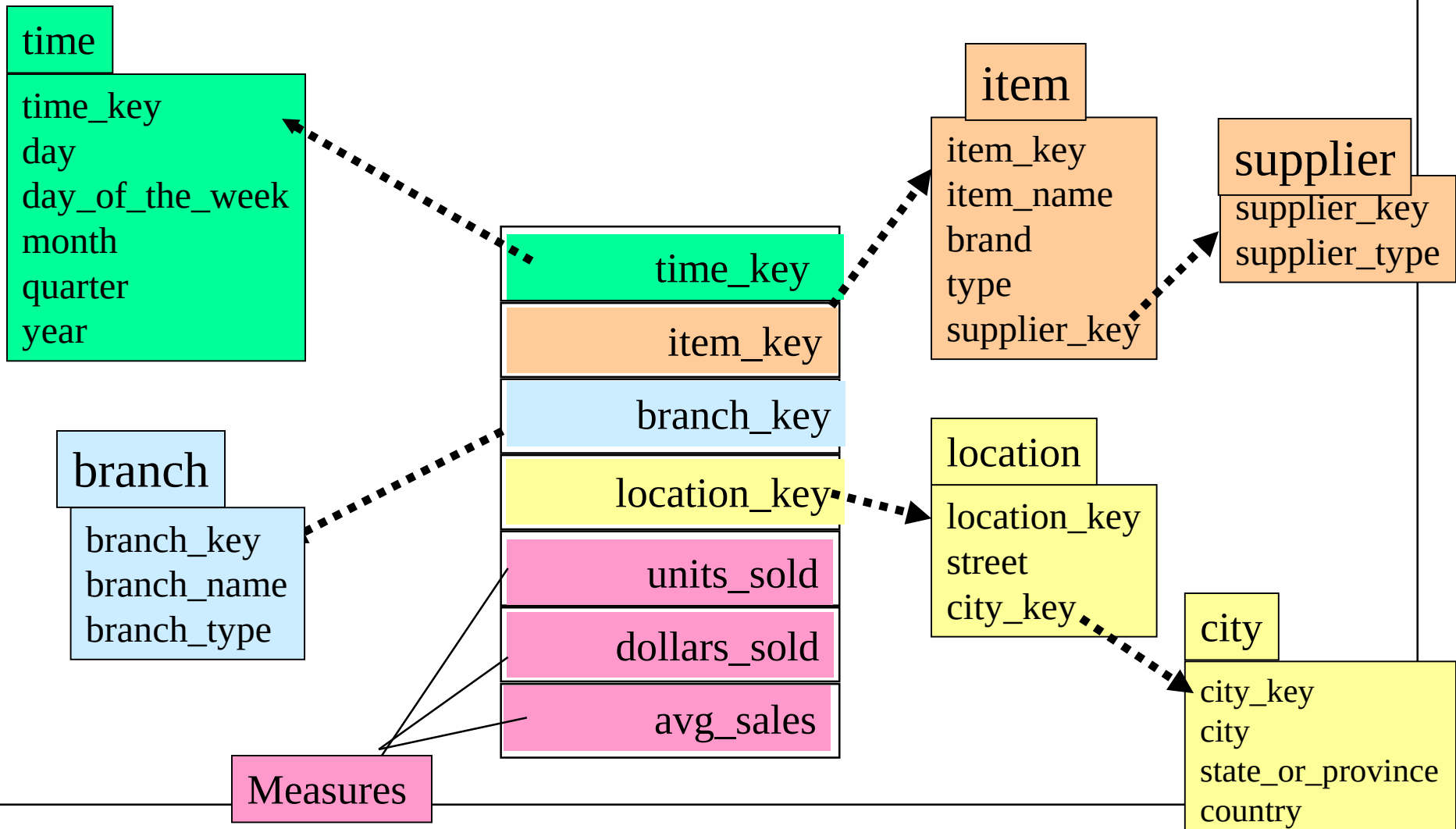
- **Example**: *product* sales in a particular *location* during a given period of *time*
- Most facts are represented by *quantitative* (numeric) values (e.g., Dollar amount of sales, number of units produced, etc.)
- Facts are stored in *fact entities*
- Quantitative values can be summed arithmetically
  - Example of *semiadditive* fact: **account balance** (can be added along the customer dimension but not along the time dimension)

Detailed Example of Star Schema

Fact entities and reference entities?



Detailed Example of Snowflake Schema



Information Granularity

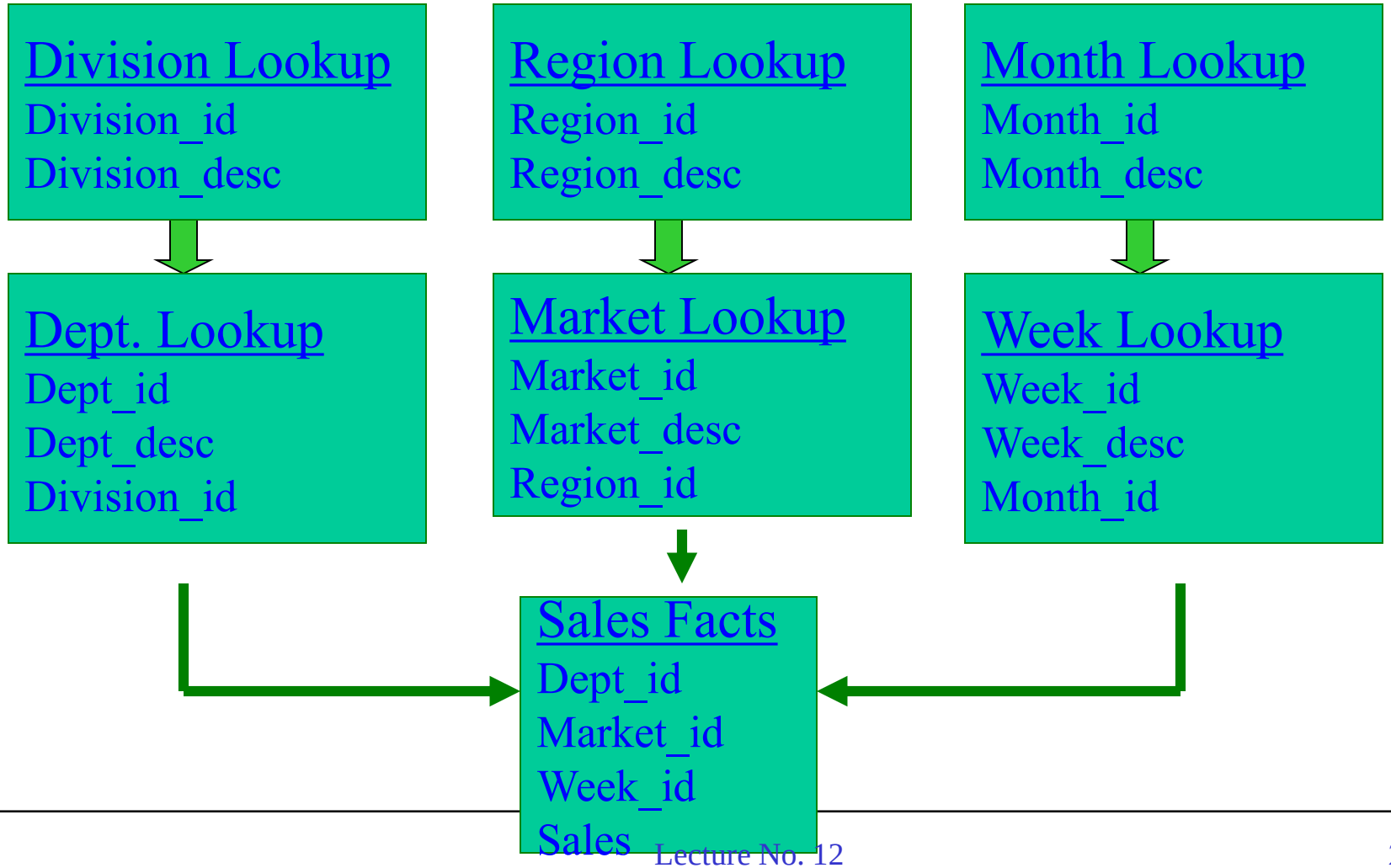
- The "optimal" level of granularity is not necessarily the lowest level of detail!
- Reducing the amount of information
  - Keep the detail data for only the minimum amount of time
  - Store only aggregated data
  - Provide both detail and summary data for a limited amount of time

DW Physical Schema

- Lookup Tables
  - Dimensional attribute data (static information)
  - Qualitative, independent data
- Relationship Tables
  - Relating attributes within a dimension
  - One-to-many relationships
  - Many-to-many relationships
- Fact Tables
  - Primary keys: attributes from multiple dimensions
  - Facts or business metrics (dynamic information)
  - Quantitative, dependent data

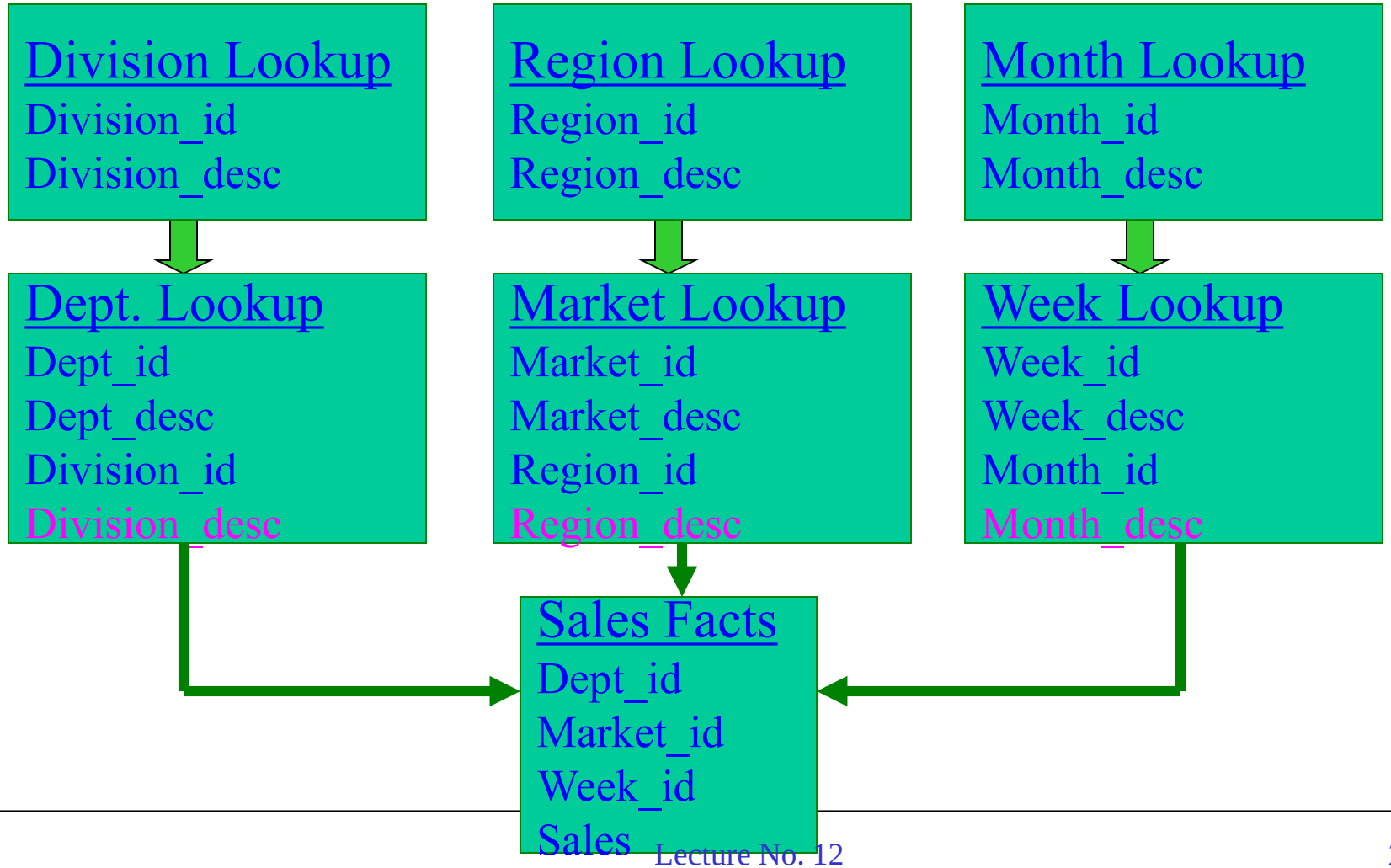
DW Physical Schema

An Example - 3NF



DW Physical Schema

An Example – De-normalization



Physical transformation of operational data

- More de-normalization
  - Store calculation results (e.g., DOB + age)
  - Enumerate measured values (e.g., file size + category)
- Attribute name matching
  - Example: WO, Job, Batch_No, Batch_ID
- Different lengths and data types
  - Example: YY vs. YYYY
- Different values for the same meaning
  - Example: M/F vs. 1/2

Physical transformation of operational data (cont'd)

- Complex data values
  - Example: Product type included in batch ID: N1002
- Missing and corrupt data
  - Using default values
  - Referencing other current data
  - Leaving the missing values as blank
  - Assigning a specific value to missing values

DW Optimization

- Aggregation / Summary Tables
  - Benefit: improved query performance
  - Drawback: increased size of the warehouse
- Partitioning
  - Time-based partitioning (helps to remove outdated data)
  - Organizational-based partitions (vulnerable to organization changes)
  - Sampling (by using random numbers)

DW Optimization (cont'd)

- Operational keys
  - Benefit: reduced transformation effort
  - Drawbacks: compound and textual keys
- Surrogate keys
  - Benefits: a layer of abstraction between DW and the source system; simple, numeric keys
  - Drawback: increased batch processing
- Indexes
  - Objective: speed-up the retrieval of records
  - Types of Indexes
    - **Primary Index**: specified on the physical ordering key field
    - **Clustering Index**: non-unique physical ordering field
    - **Secondary Index**: specified on any non-ordering field

DW Project - Practical Example

Manpower Efficiency System

Objective: calculating average efficiency in every production department (work cell)

$$\text{Average Efficiency} = \frac{\text{Total Standard Time}}{\text{Total Paid Time}}$$

Manpower Efficiency System

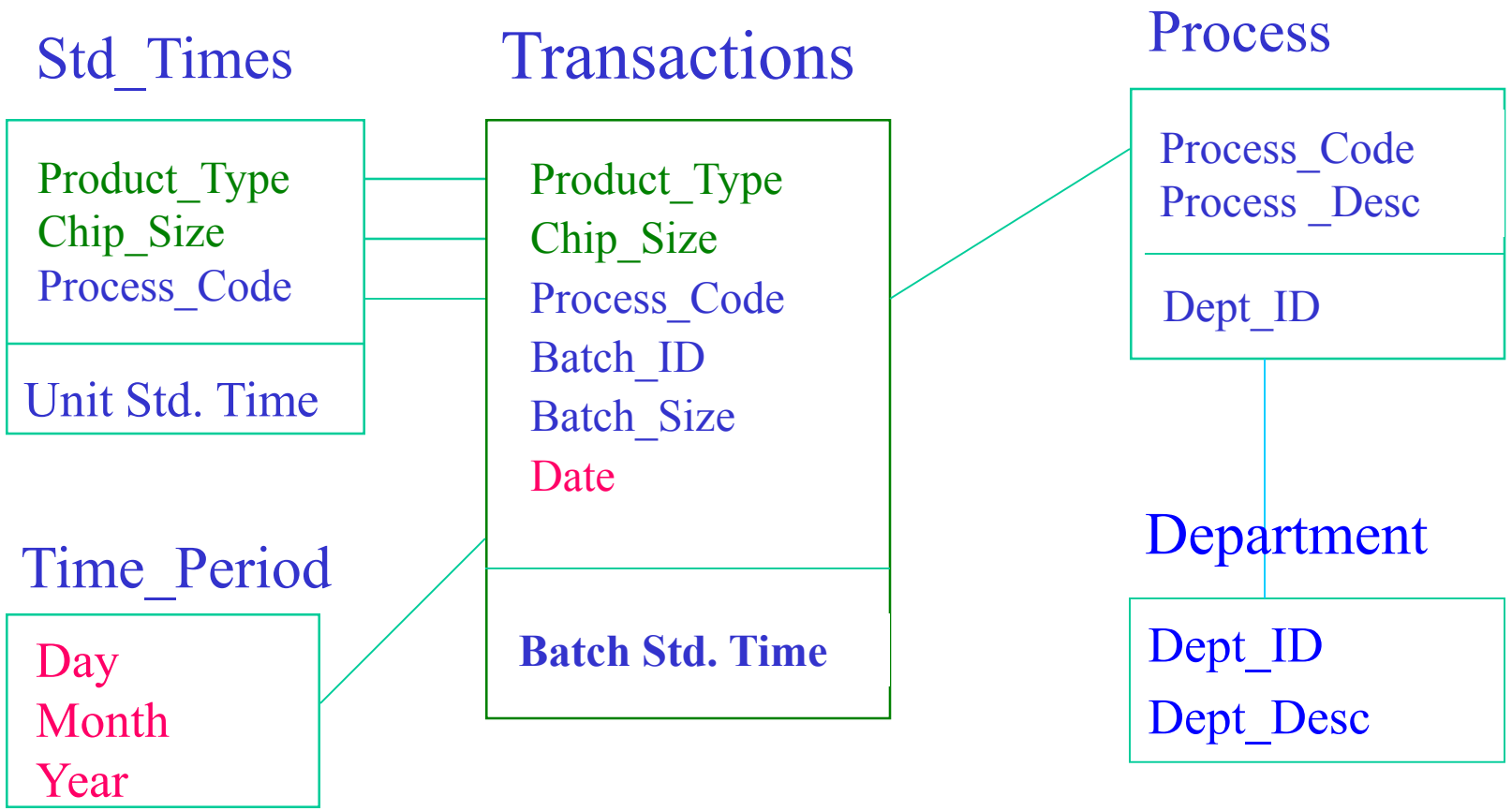
Why Data Warehouse ?

- Average Amount of Transactions: **700** per day
- Number of Records in Std_Times Table: **400**
- Amount of Table Joins Required to calculate standard time of execution for every transaction (a Cartesian Product):
  - **280,000** per day
  - **1,400,000** per week
  - **6,160,000** per month

Conclusion: Efficiency calculations can be very inefficient !

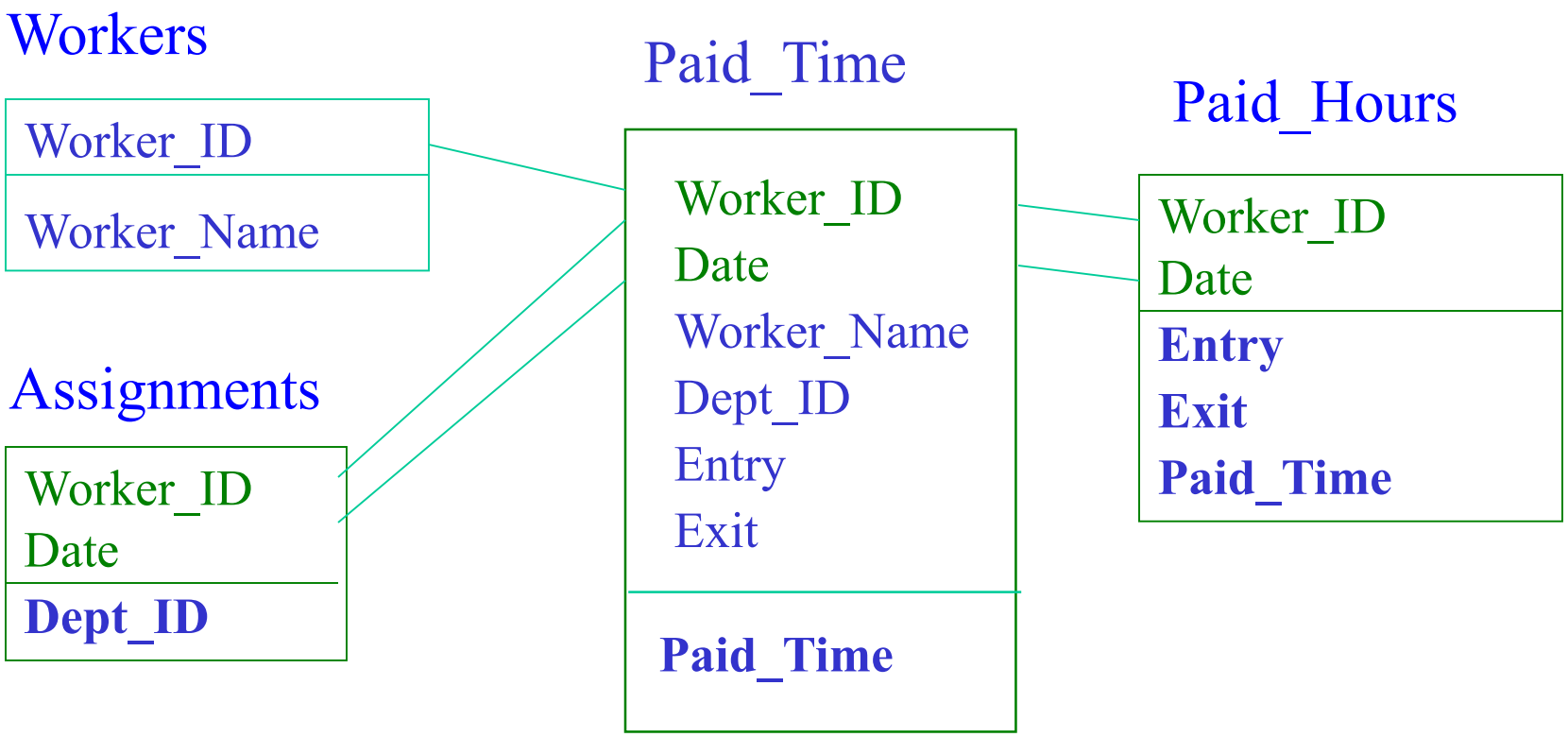
Manpower Efficiency System

Star Schema - Transactions



Manpower Efficiency System

Star Schema - Paid Hours



Manpower Efficiency System

Surrogate Keys

Assignments (OLTP)

Worker_ID:105

Date: 01/10/95 (34973)

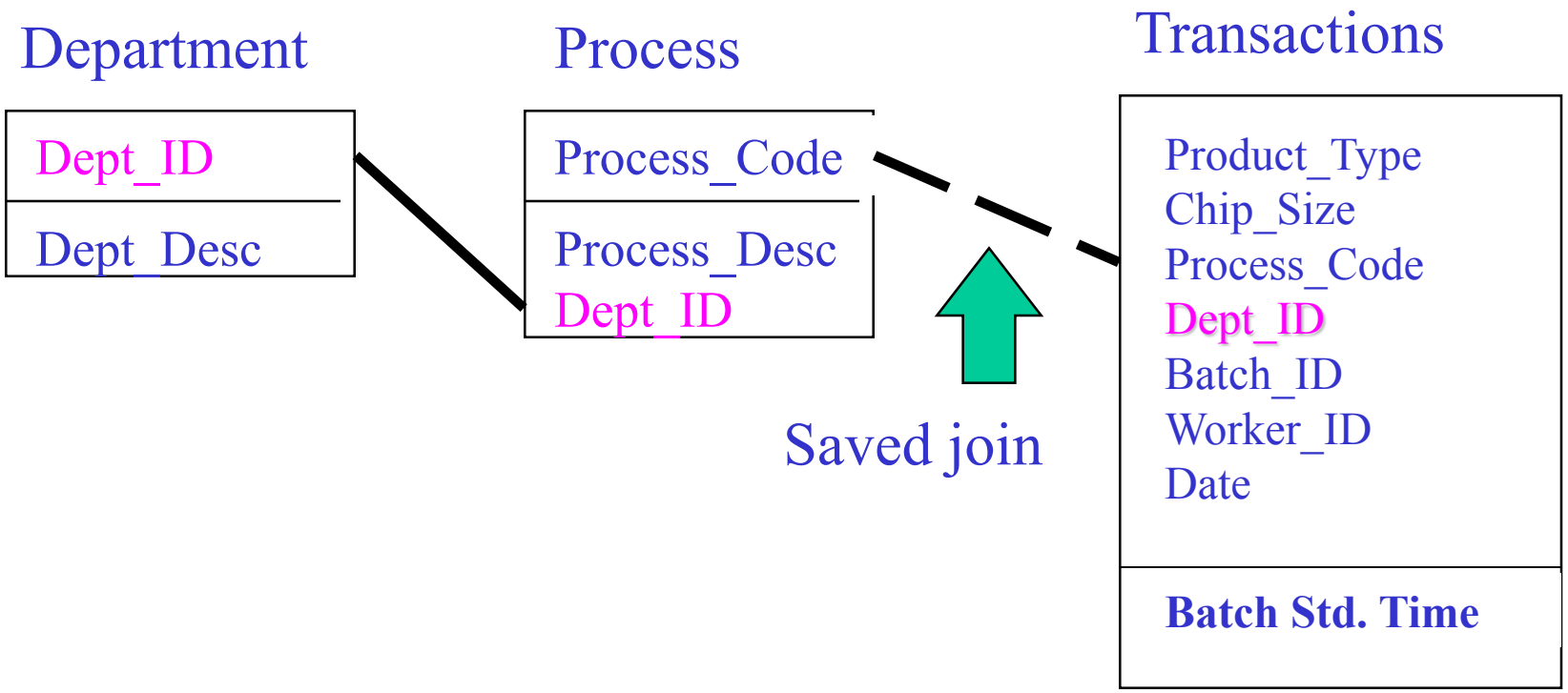
Dept_ID

Paid_Time (DW)

Key: 34973.105

Manpower Efficiency System

Denormalization of Tables



Manpower Efficiency System

Partition and Aggregation

One week

Six Months

Transactions

Product_Type Chip_Size Process_Code Dept_ID Batch_ID Worker_ID Date
Std. Time

Paid_Time

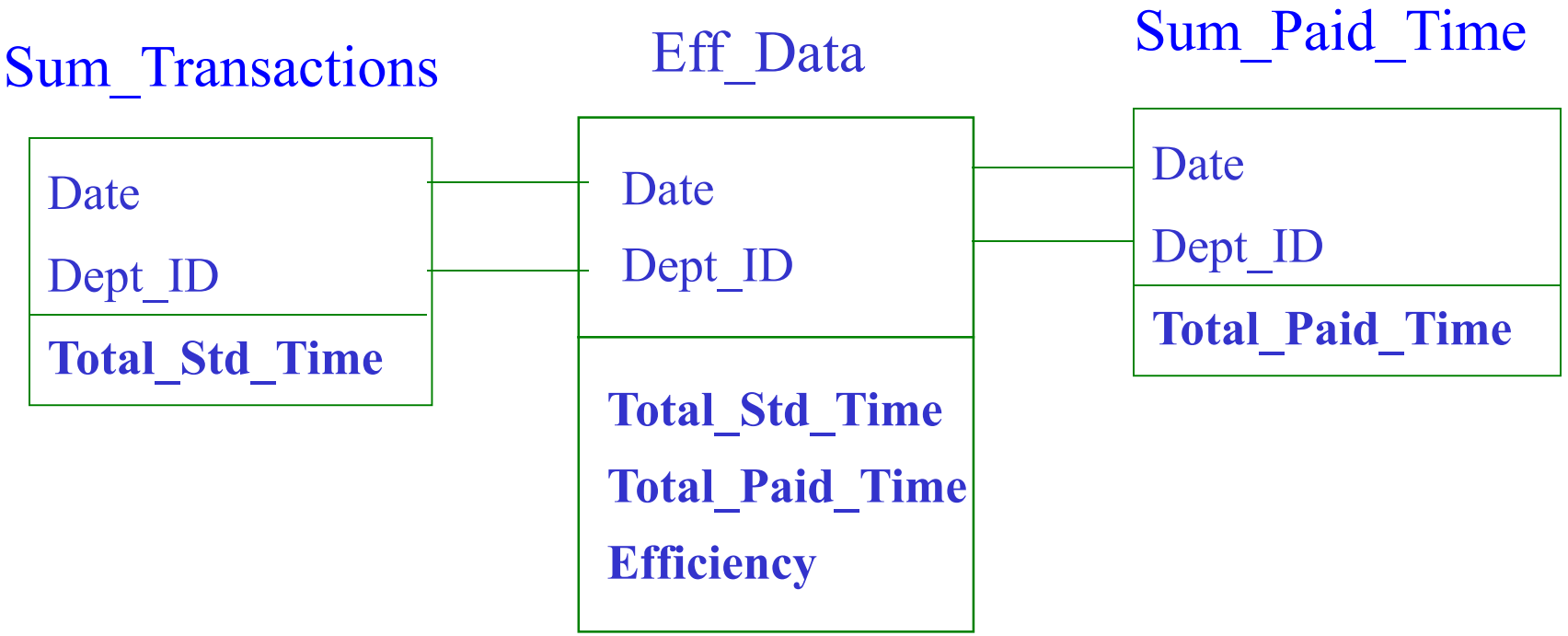
Worker_ID Worker_Name Dept_ID Date Entry Exit
Paid_Time

Eff_Data

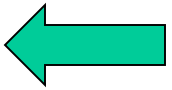
Date Dept_ID
Total_Std_Time **Total_Paid_Time** **Efficiency**

Manpower Efficiency System

Star Schema - Aggregated Data



Business Intelligence and Data Warehousing

- Business Intelligence
- Why Data Warehousing?
- Data Modeling
- Metadata ("data about data") 
- Data Quality
- ETL (Extraction, Transformation, and Loading)

Metadata - Definitions

- Metadata is "**data about data**" - *Inmon, 1994*
- Metadata is **high-level data that describes lower-level data** - *APT Data Group 1996*
- Considering that we don't know exactly what it is, or where it is, we spend more time talking about it, worrying about it, and feeling guilty we aren't doing anything about it than any other topic - *Kimball, 1998*
- A **map** to the data in the data warehouse – *Sperley, 1999*

Data without Metadata

Examples

- Country = 972
- City = 068
- Department = 372
- Date = 07/04/98
- Gender = 0
- Age = 370

Data with Metadata

Examples

- Country = 972 (Israel)
- City = 068 (Tel-Aviv)
- Department = 372 (Information Systems Engineering)
- Date = 07/04/98 (4 July 1998)
- Gender = 0 (Female)
- Age = 370 (70 years)

Sources of Existing Metadata

- Code Documentation
- OLTP Applications
- DBMS (Database Management System)
- Middleware (Extraction Software)
- DW Design Documentation
- CASE Tools
- User Tools

Classification of Metadata

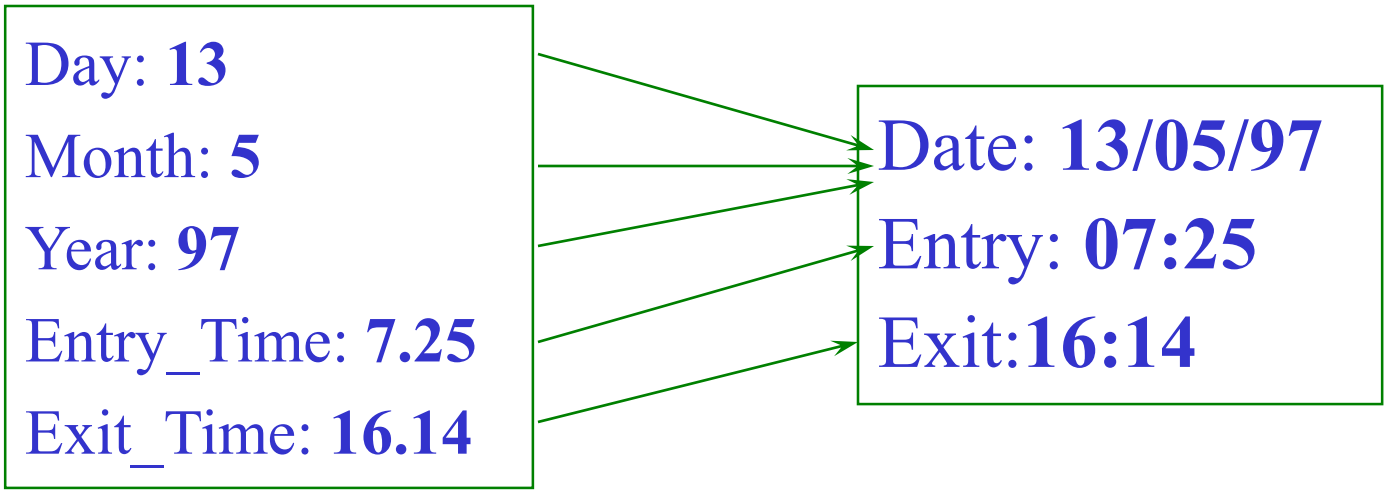
- Implementation-time metadata
  - Entities, locations, and motivations
- Run-time metadata
  - Active metadata
    - Manages security and access to data
    - Provides usage data
  - Passive metadata
    - Creates the **context** for the business data

Manpower Efficiency System

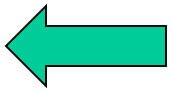
Data Transformation and Metadata

Paid_Hours (OLTP)

Paid_Time (DW)



Business Intelligence and Data Warehousing

- Business Intelligence
- Why Data Warehousing?
- Data Modeling
- Metadata ("data about data")
- Data Quality 
- ETL (Extraction, Transformation, and Loading)

Causes of Poor Data Quality

- Process Problems
  - Data entered at the wrong point in the operational process
  - Inaccurate measuring and counting equipment
- People Problems
  - Failing to update the data on time
  - Entering incorrect data by mistake ("keying errors")
  - Entering incorrect data on purpose (fraud)

Examples of Data Quality Problems

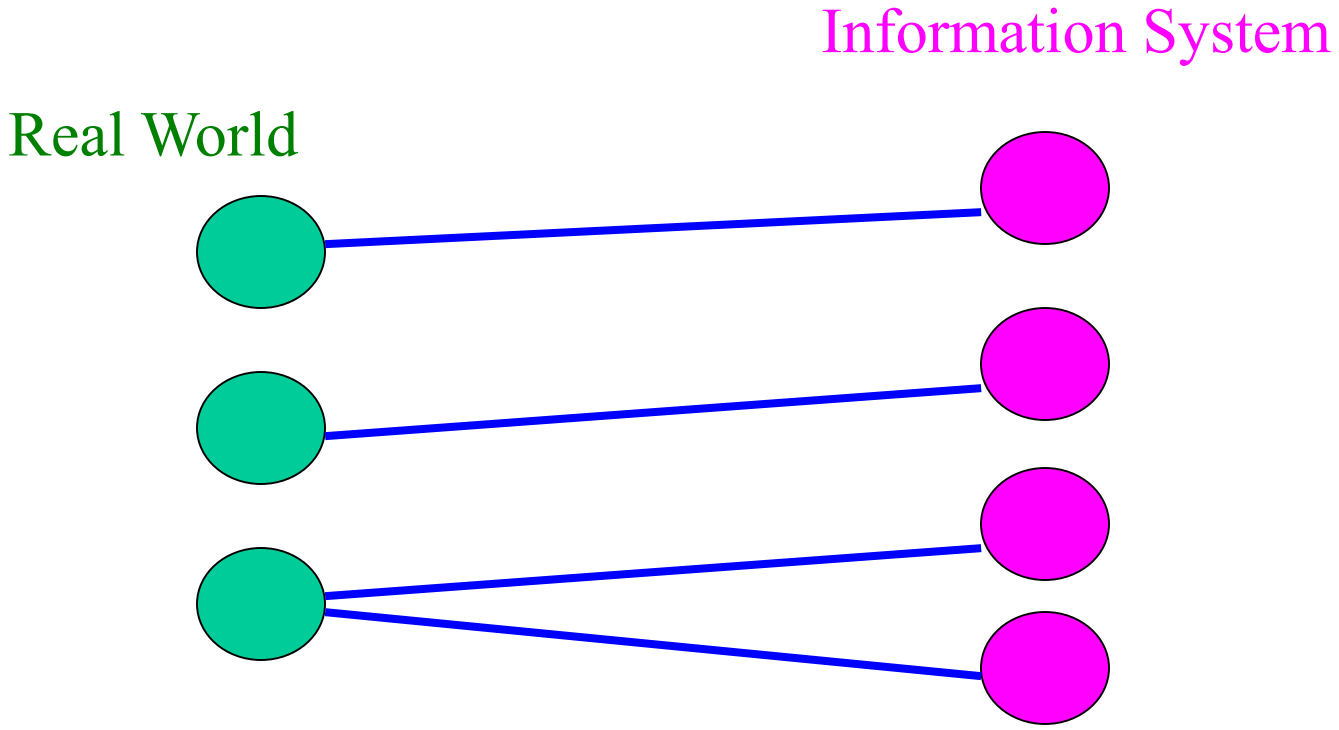
Source: http://www.bcs.org/upload/pdf/ewrazen-120607.pdf

- Retail company found over 1m records contained home tel number of "000000000" and addresses containing flight numbers
- Insurance company found customer records with 99/99/99 in creation date field of policy
- Car rental company discovered duplicate agreement numbers in their European data warehouse
- Healthcare company found 9 different values in gender field
- Food/Beverage retail chain found the same product was their No 1 and No 2 best sellers across their business

Data Quality Dimensions

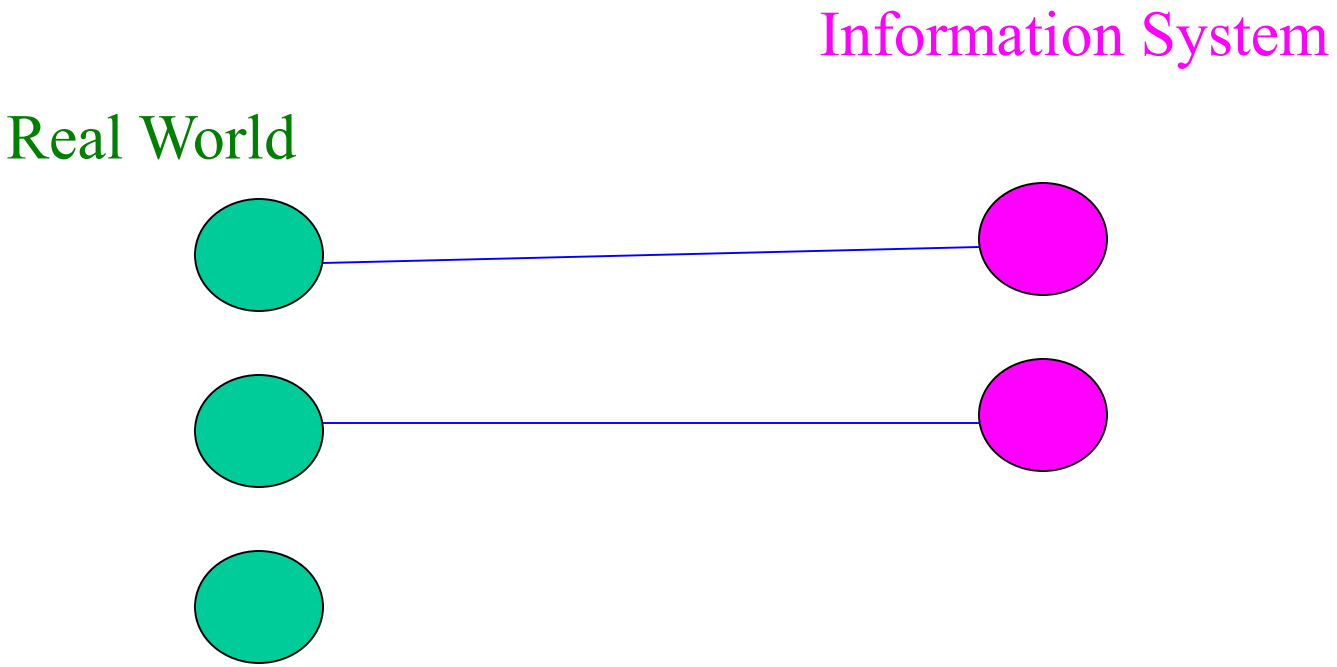
(based upon Wand and Wang, *Comm. of the ACM*, Nov. 1996)

1. Proper Presentation of Data



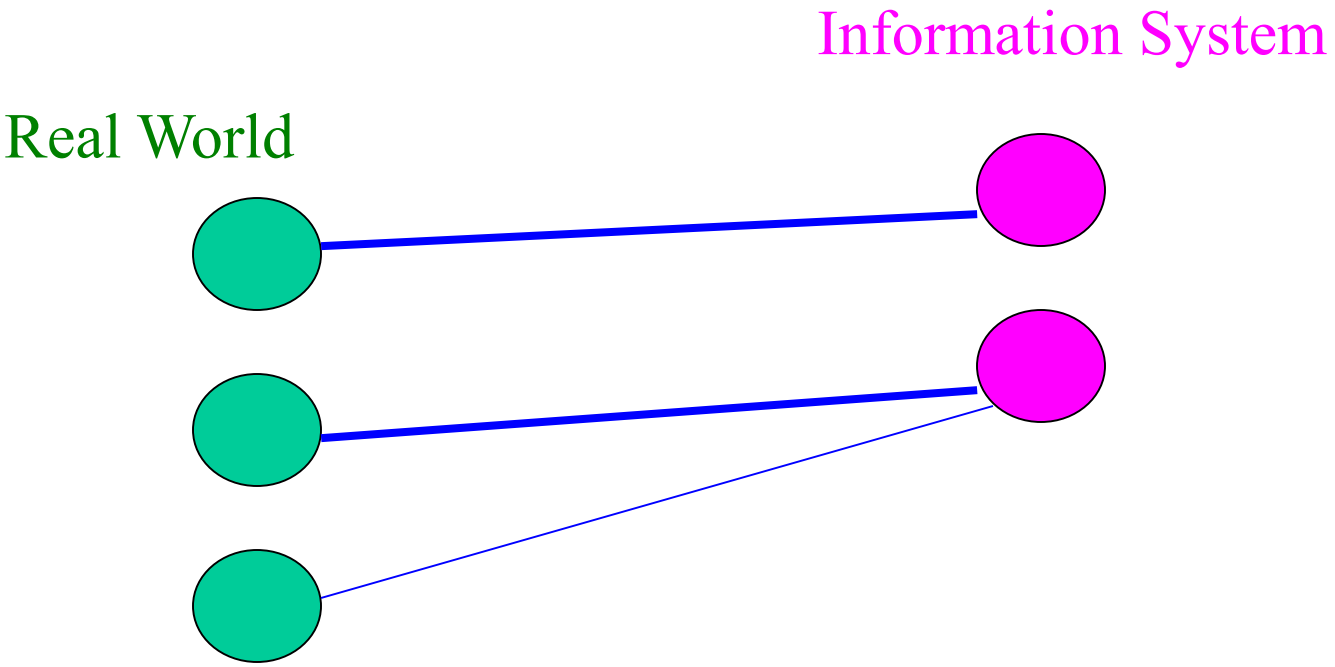
Data Deficiency

2. Incomplete Presentation



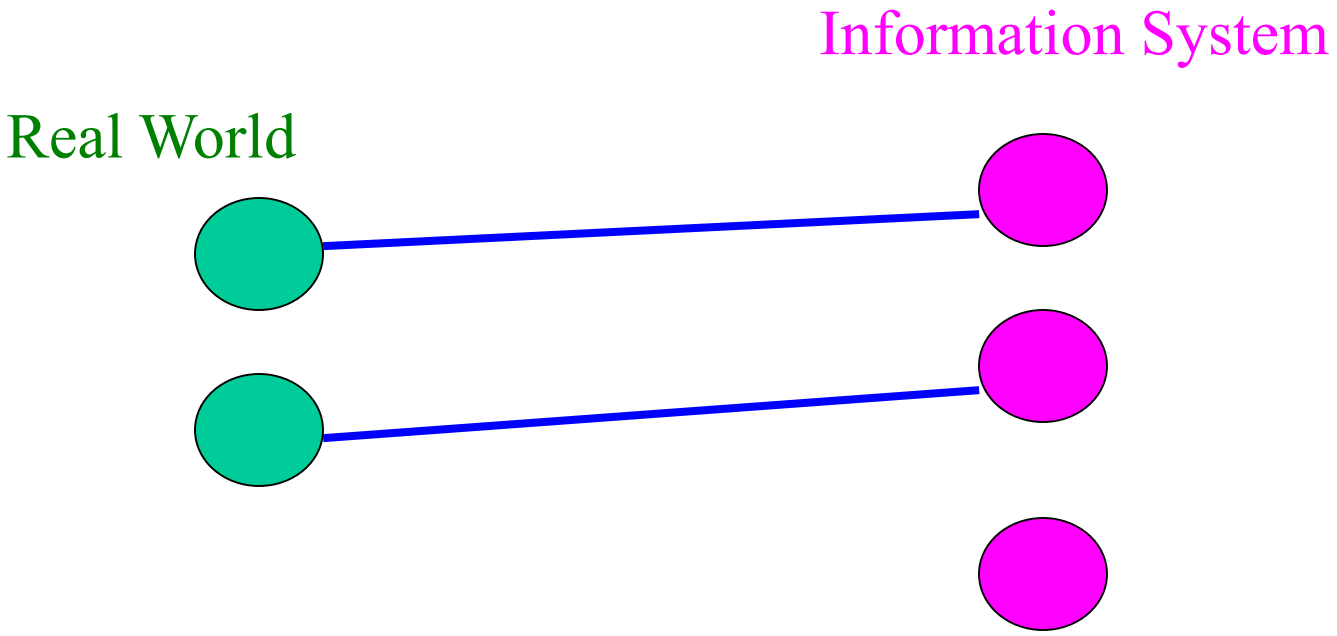
Data Deficiency (cont.)

3. Ambiguous Presentation



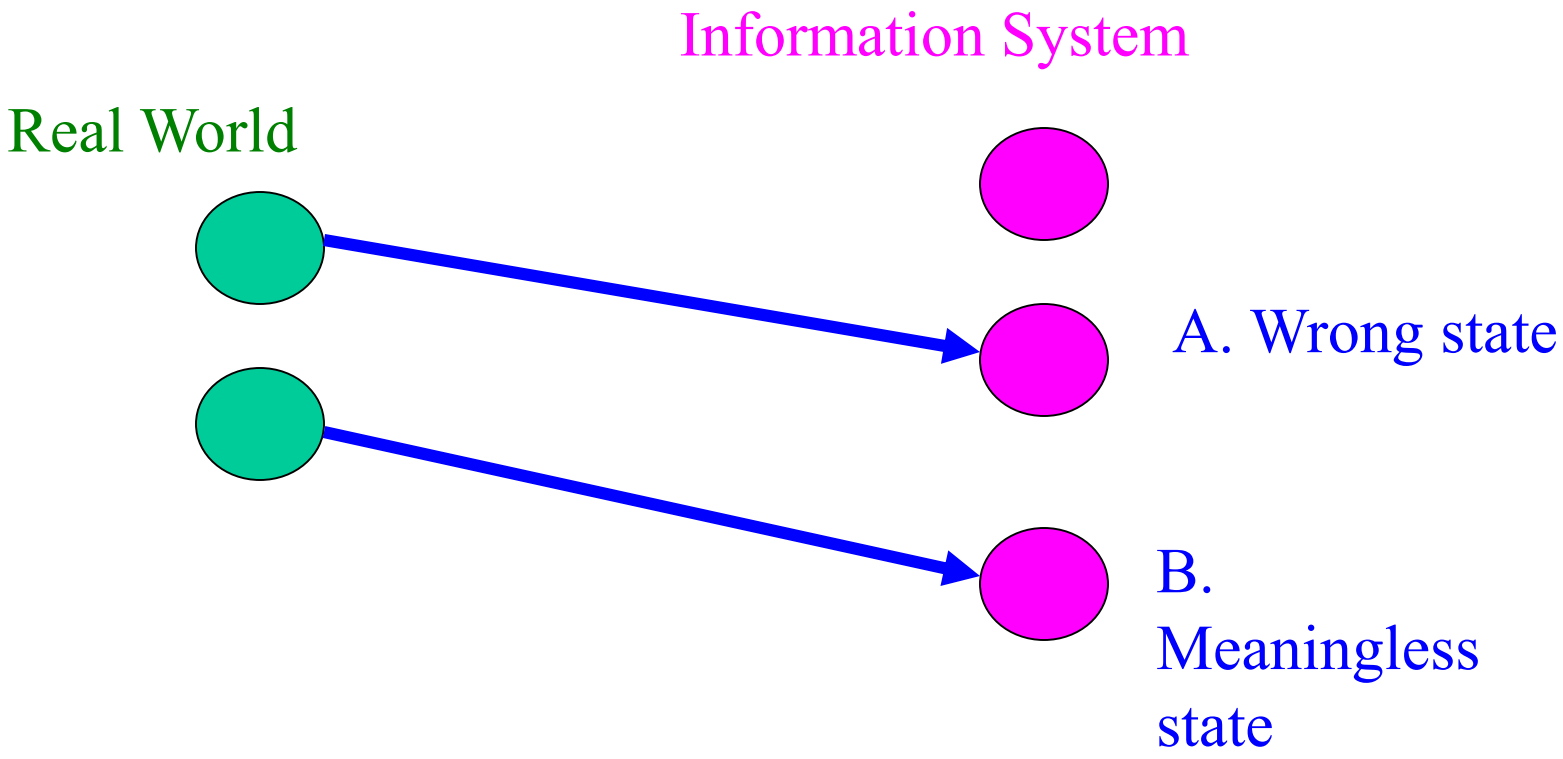
Data Deficiency (cont.)

4. Meaningless State



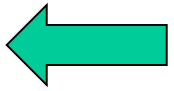
Data Deficiency (cont.)

5. Data Garbling



Business Intelligence and Data Warehousing

- Business Intelligence
- Why Data Warehousing?
- Data Modeling
- Metadata ("data about data")
- Data Quality
- ETL (Extraction, Transformation, and Loading)



ETL – The Words Behind the Acronym

- Extraction
  - Extract data from a source database
- Transformation
  - Transform the data into a format suitable for a target database
- Loading
  - Load the data into the target database
- Common Practice
  - ETL requires about **80%** of the efforts of building a DW

Steps of DW Loading

(based upon Ralph Kimball, *Mastering Data Extraction*, DBMS and Internet Systems, June 1996**)**

1. Read the legacy data
   - choosing data repository
   - physical transformation of data
   - using metadata
2. Decide what changed
   - new facts
   - changes in dimension tables
   - using record timestamps

Steps of DW Loading (cont'd)

3. Generating keys for changing dimensions
   - objective: track dimension changes
4. De-normalization of dimensions
   - objective: combining separate sources into single records
5. Create load record images
   - only detail data
   - no generated or aggregate (summary) records

Steps of DW Loading (cont'd)

6. Migrate the data from the Operational system to Data Warehouse
7. Create aggregate records
8. Generate artificial keys for aggregate records
9. Bulk load all the records
   - enforce referential integrity (star schema)

Steps of DW Loading (cont'd)

10. Process load exceptions
    - records failing the referential integrity check
11. Index the newly loaded records
12. Quality assurance
    - data cleaning
    - compare totals to the operational data
13. Publish the data
    - email to all users

Manpower Efficiency System
Data Loading

- Transactions Table:
  - Loading operational data for one day or a number of days
  - New data is appended to existing table
- Paid_Time Table:
  - Loading operational data for a full or a partial week (starting with Sunday)
  - New data overwrites the existing data on the same week

Lecture No. 2 - Summary

- DWH is a *subject-oriented*, *integrated*, *time-variant*, and *non-volatile* collection of data in support of management's decision making
- DWH and OLTP are *different* in many aspects
- A DWH stores *dimensions* and *facts*
- It is important to *optimize* the DWH performance
- Metadata is a *map* to the data in the DWH
- Data quality is measured by various *quality dimensions*
- The ETL process has 13 stages