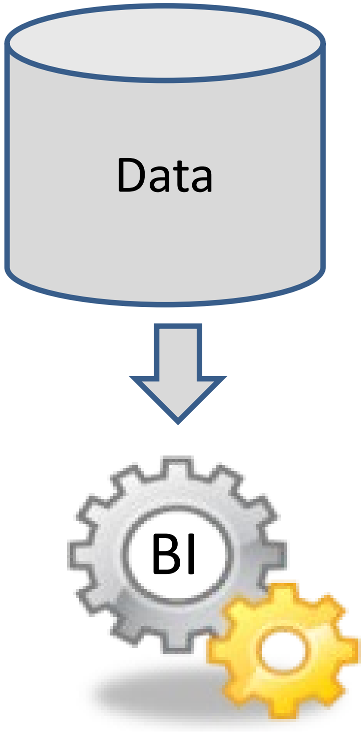
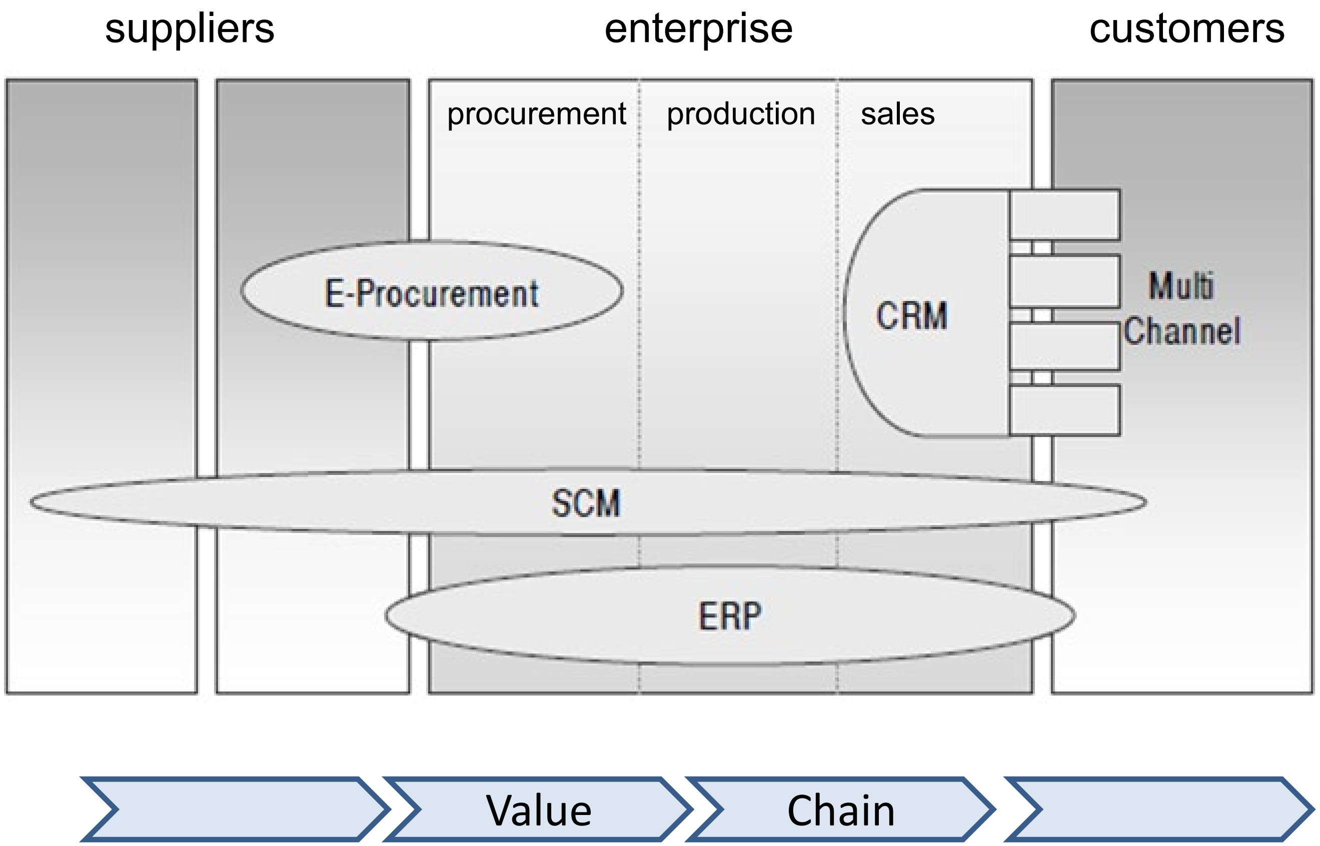
BI-Tools Backend: Data Warehouse

Knut Hinkelmann

- … transform **raw data** into meaningful and useful **information**…
- **Raw data** is the starting point!



Where the data come from (1)



CRM – Customer Relationship Management
SCM – Suppy Chain Management
ERP – Enterprise Resource Planning

adapted from Kemper et al. 2004

Where the data comes from (2)

- Internal data sources:
  - (Transactional) standard business applications: sales data, accounting, SCM, ERP, CRM, …
  - Legacy databases, spreadsheets
  - Web data: clickstreams from server logs, application logs
  - textual documents (from DMS, CMS, intranet, email,…)

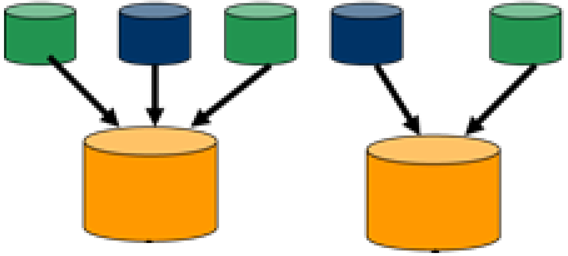
structured-ness

- External data sources:
  - Web and web 2.0

CRM – Customer Relationship Management
SCM – Suppy Chain Management
ERP – Enterprise Resource Planning

BI tools – backend

- Observations:
  - many questions involve multiple (types of) data
  - sometimes the data can be expected to originate from more than one source system
  - for answering the questions, data from various sources needs to be connected
    - example: «Which is the best way to distribute product XYZ to customers?” → involves information about customers (e.g. profitability, behaviour) as well as about channels (e.g. cost of each channel)



Planning Data vs. Operative Data (1)

- **operative data:** generated by and used in processing operational transactions (on-line transaction processing, OLTP)
  - many concurrent users access and modify the same data
  - focus on transactions
  - example: booking/reservation systems

- **planning data:** used for decision support
  - read-only data

following Kemper et al. ch 2.1

Planning Data vs. Operative Data(2)

Different requirements for management planning data and operative data

	Operative data	**Planning data**
users	clerk IT professional	knowledge worker decision maker
function/goal	support **day to day operations**	**decision support**
DB design	application-oriented	subject-oriented
data	current, **up-to-date** detailed information on business events, flat relational	**historical**, summarized, multidimensional integrated, consolidated
usage	Continuous, repetitive, concurrent	ad-hoc
access	**read/write**	**read-only**
queries	Static, transactions embedded in application code	Ad-hoc, for changing information needs
metric	transaction throughput	query throughput, response

Separate Management of Planning Data → Data Warehouse

adapted from http://www.slideshare.net/idnats/data-warehousing-and-data-mining-presentation-725476

Data warehouse

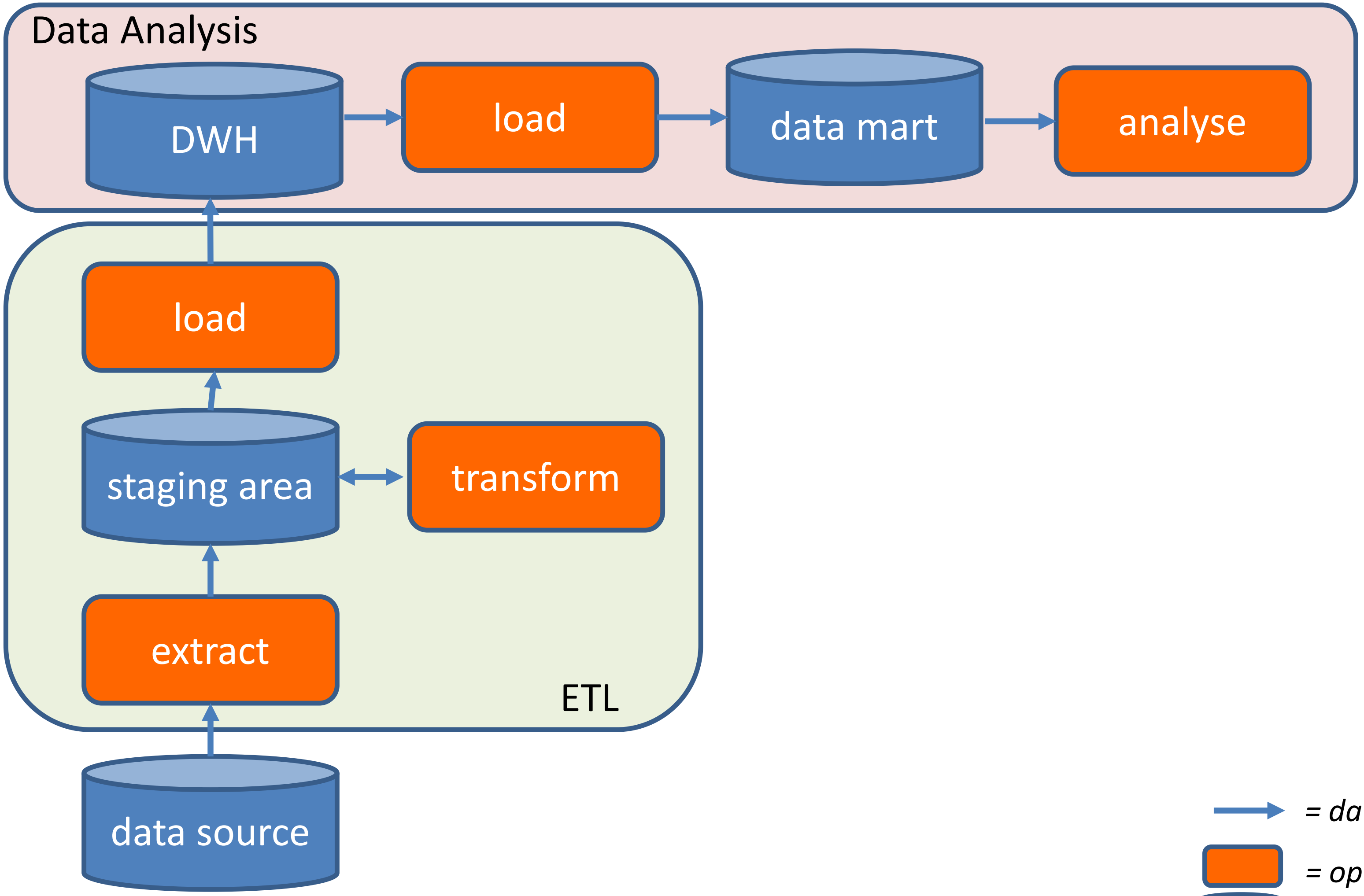
- *A data warehouse is*
  - *“a **copy** of transaction data specifically structured **for querying and reporting” (Kimball et al. 2008)***

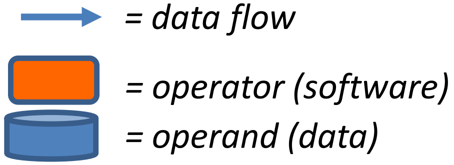
  or

  - *“an environment […] comprising a data store and […] tools for data extraction, loading, storage, access, query and reporting […] to **support decision-oriented management queries**” (Bashein/Markus, 2000)*

- *Data warehousing is the entire process of data **extraction, transformation, and loading** of data to the warehouse and the **analysing** the data by users and applications.*

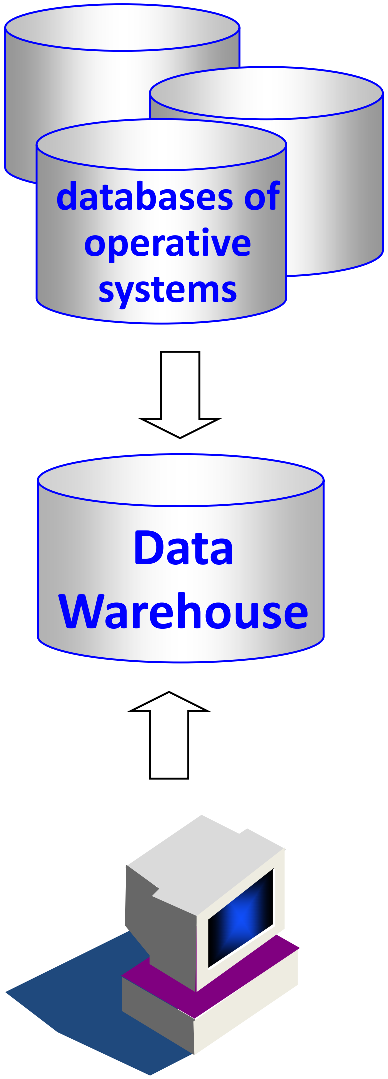
Reference Architecture – Overview





based on Bauer/Günzel, fig. 2-1

Data Warehouse



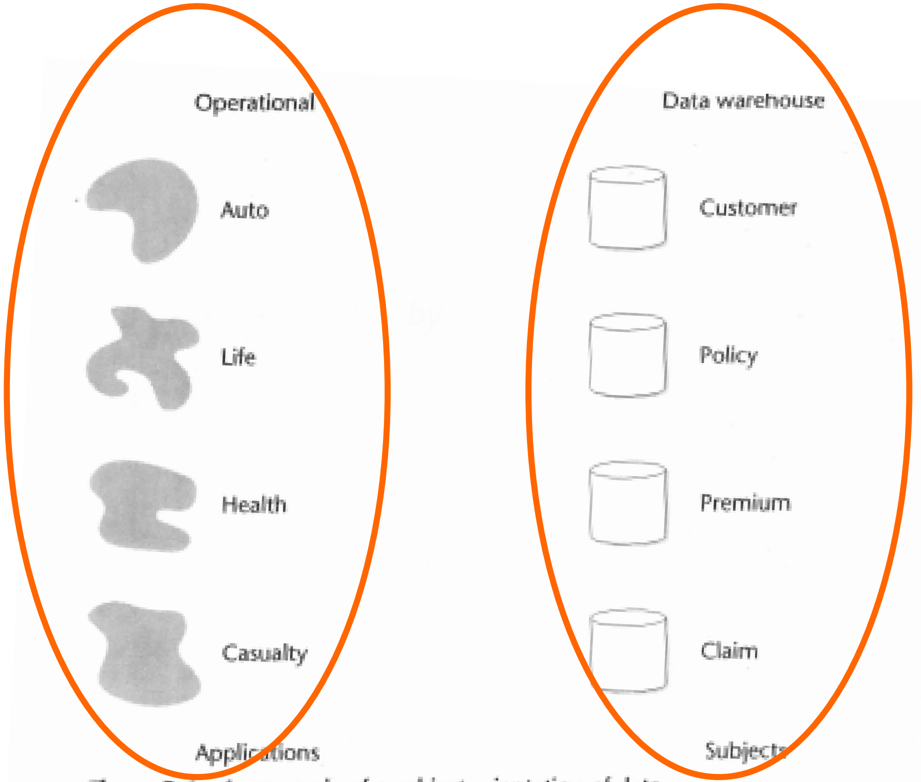
- A Data Warehouse is a database that supports strategic decisions by providing
  - high-volume and regular excerpts from operative databases
  - often aggregated[1]
  - also for ad hoc[2] analysis
- Essential characteristics (Inmon 2005):
  - Subject oriented
  - Integrated
  - Time variant
  - Nonvolatile

[1] combined, consolidated (e.g. als sum, average, indicators)
[2] without preparation, in contrast to standardized analysis

Subject orientation

- Example: insurance

operational systems are organised around functional applications



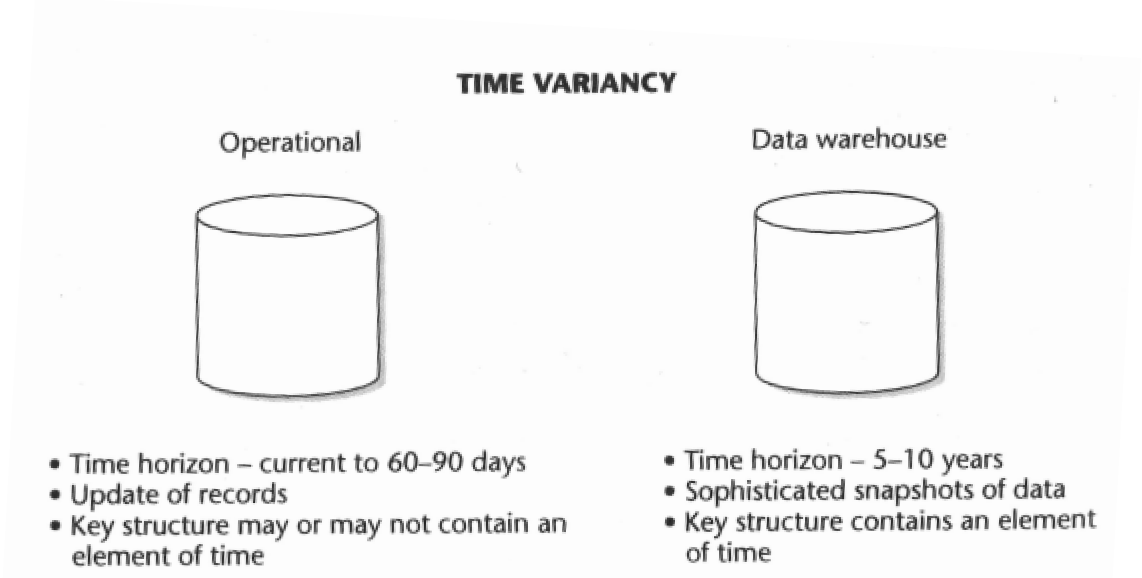
Fīeure 2-1 An example of a subject orientation of data.

analytical systems should be organised around the major subject areas of an insurance!

figure taken from B. Inmon: Building the data warehouse.

Integration, Time variance

- **Integration:** provide a «single version of the truth»
  - remove redundancy, inconsistency, semantic contradictions (see ETL processes later)
- **Time variance:** DWH maintains historic data, data is collected over a long time period with information when it was valid



- Time horizon – current to 60–90 days
- Update of records
- Key structure may or may not contain an element of time

- Time horizon – 5–10 years
- Sophisticated snapshots of data
- Key structure contains an element of time

Figure 2-4 The issue of time variancy.

figure taken from B. Inmon: Building the data warehouse.

Non-volatility

- **Non-volatility:** data is not updated by end users on a regular basis
  - bulk loading, «read-only» access

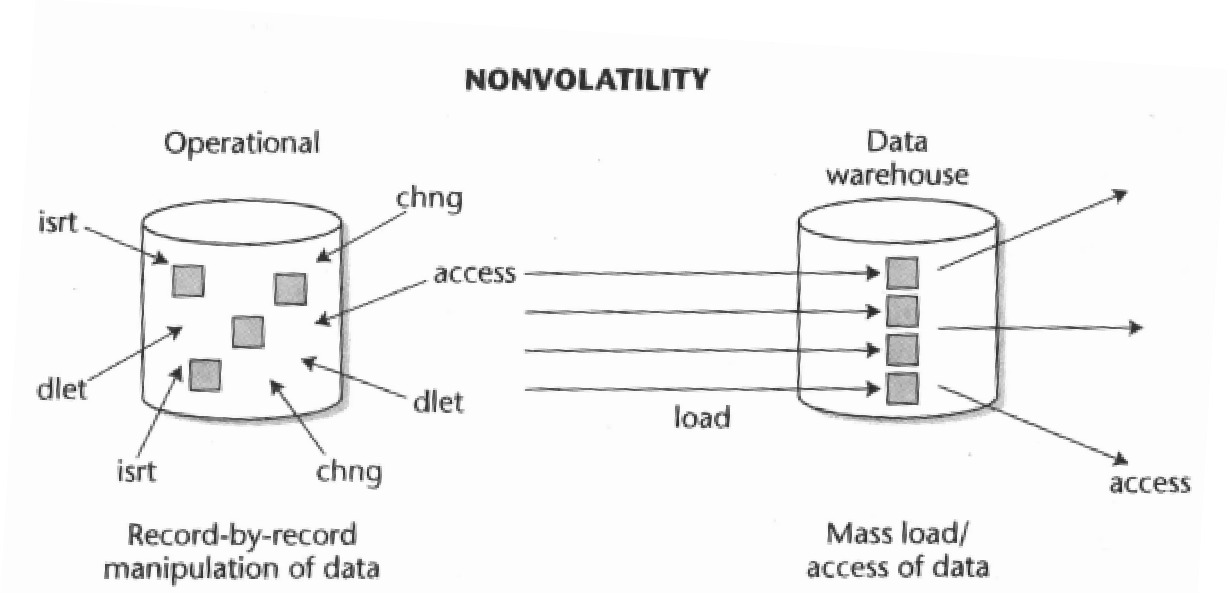


Figure 2-3 The issue of nonvolatility.

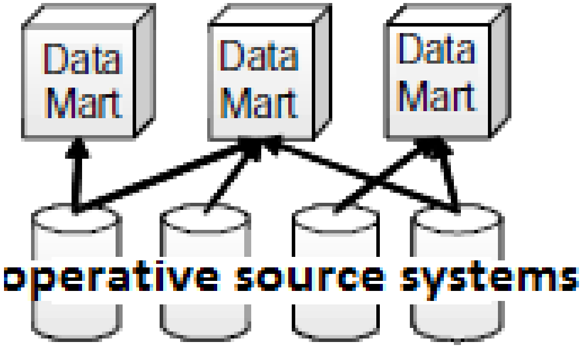
figure taken from B. Inmon: Building the data warehouse.

Data Marts

A **data mart** stores data for a *limited* number of subject areas.
It is used to support *specific* applications.

Independent data mart

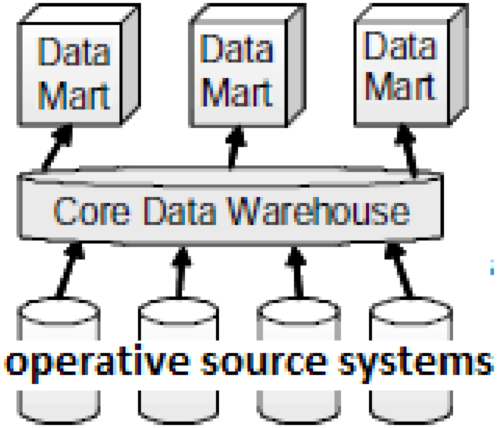
- created directly from source systems
- Possibly joined into a data warehouse later



following Kemper et al. fig. 2.4

Dependent data mart

- Source data are aggregated into a data warehouse
- data marts are created as subsets (e.g. for efficiency reasons)

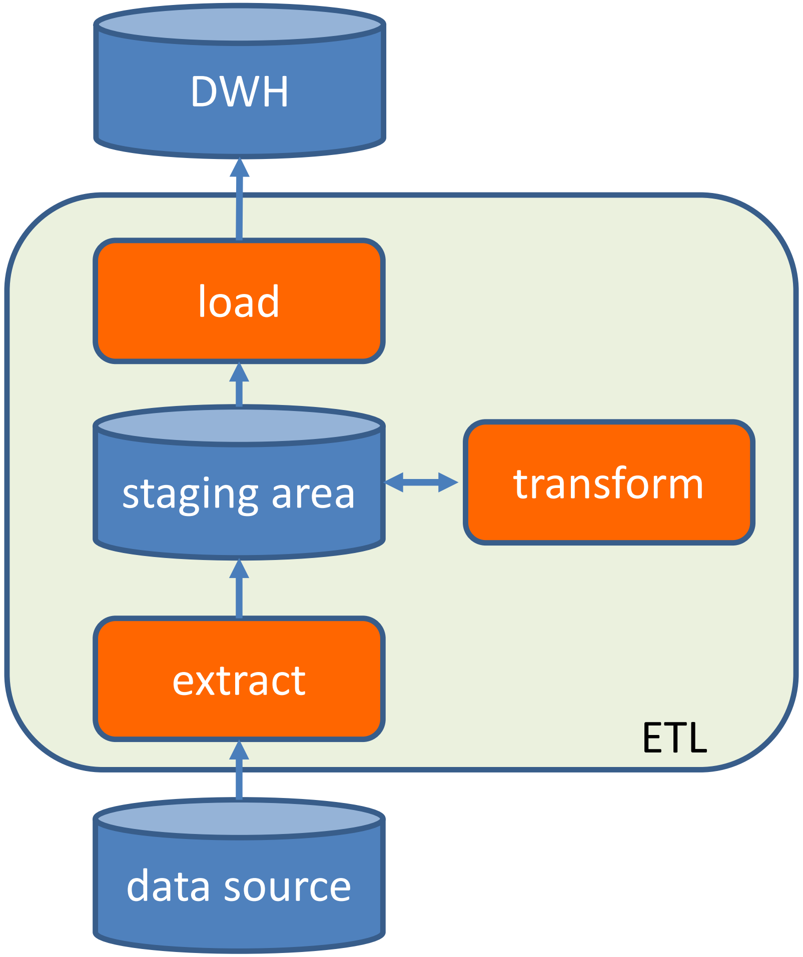


Data Marts: Departmental vs. Enterprise?

- Question: should data marts be enterprise-wide or departmental?
  - Answer 1: Data marts should be organised around business processes (orders, invoices,…), not department boundaries!
  - Answer 2: … but they don't necessarily have to be enterprise-wide (depends on the business process)!

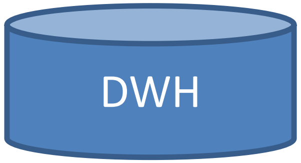
ETL – Extract,Transform, Load

ETL process



- The process of
  - ♦ **extracting** relevant data from source systems
  - ♦ **transforming** the data into the target format defined for the DWH or data mart
  - ♦ **loading** the data into the DWH

Reference architecture – DWH



- DWH = integrated data basis for all analyses
  - integration means both schema and data integration of various sources => «single point of truth» (no by-passes allowed!)
  - purpose: flexibility for re-using the data in multiple anlyses, no focus on (and hence no pre-aggregations for) particular types of questions
  - provides a collection of data that can be used to build data marts for specific analyses
  - not always present because of high cost: building data marts for specific analysis purposes directly from source data is often cheaper…

extract

ETL process – Extraction

- **Extraction =** Collecting information to be added into DWH
  - Read source data into staging area
  - Control the selection of data that should be copied
- Extractions can happen…

  …periodically

  …on human request

  …event-based (e.g. when a certain number of changes has occurred)

  …upon each change

Reference architecture – staging area

staging area

- A copy of the source data is stored in the staging area
  - one-to-one image of the source!
  - no relations, no integration (yet)
- Then, transformations are run on this copy – without impact on the operation of both the data sources and the DWH
- After completion of all necessary transformations, data is copied to the DWH and deleted from the staging area.

ETL process – Transformation

- **Transformation =** adapting data, data quality and schemas to the requirements of users
  - **Filtering:** remove syntactic and semantic defects of data
  - **Harmonisation:** map source schemas to the target schema of the DWH
    - syntactic harmonisation: schema integration + data integration
    - business harmonisation
  - **Aggregation:** aggregate data along dimension hierarchies (e.g. «customer», «customer segment», «all»)
  - **Enrichment:** pre-compute values of frequent interest and store as new attributes
    - on the basis of harmonised/aggregated data

ETL- Filtering: Error Classes

	1. class: Automatic identification	2. class: (Semi-) automatic identification
Syntactic	- Known formatting variants (abbreviations, date formatting etc.) - encoding problems	- Spelling variants/errors
Semantic	- Missing values (incompleteness) - redundancy (duplicates) - non-unique identifiers - missing referential integrity	- Incorrectness (e.g. outliers) - inconsistencies (violating business rules or contradictions) - dummy values

adapted from Kemper et al. fig 2.8

Exercise Data Integration

- Consider the following rows from a database table with customer information:

CustId	Customer Name	Customer Category	Contact Person	First contact	Most recent contact
23435	Univ. of Pennsylvania	Academic	Michael Gordon	21/01/2002	11/10/2011
87394	FunIT Corp.	Industry	Ian Finnegan	03/07/2008	10/01/2012
87394	Telly Inc.	Services	Susan Smith	14/10/2011	01/09/2011
16572	Uinversity of Pennsylvania	Academic	Michael Gordon	21/01/2002	11/10/2011

- Which of the following categories can be found in the table?

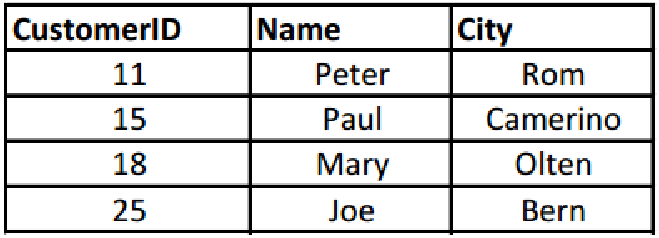
☐ incompleteness
☐ duplicate records
☐ non-unique ids
☐ spelling errors
☐ inconsistencies
☐ incorrectness

ETL- Filtering: Correction Measures

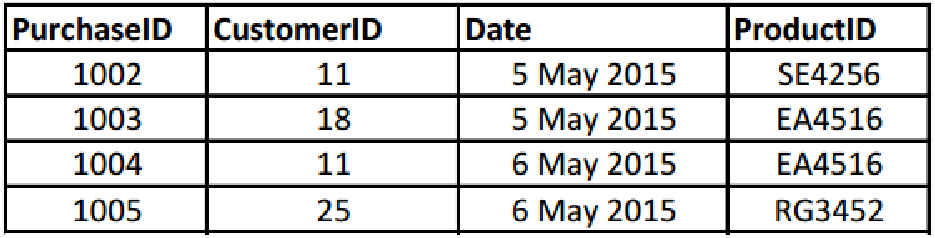
- Correction measures
  - **1st class:**
    - *incompleteness:* define rules to fill in missing values (e.g. fill sales values with ones from previous month or planned ones)
    - *duplicate detection:* often there is a combination of values that unambiguously identifies a record => if these are the same, match!
    - *formatting/encoding/non-unique id issues*: simple scripting
  - **2nd class:**
    - *spelling variants/errors:* use string similarity, thesauri (extend as you go along)
    - *general incorrectness: hard to spot automatically, can define automatic sanity checks…*
    - *outliers:* statistic analyses
    - *inconsistencies:* checks based on business rules

ETL - Harmonization

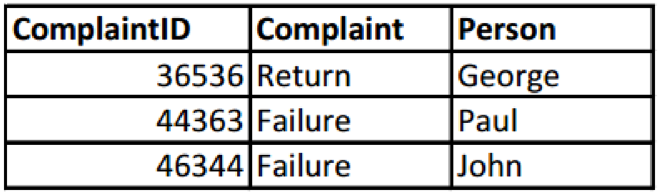
- These are parts of tables that should be integrated in a DWH. What harmonisation tasks/problems do you see?



CustomerID	Name	City
11	Peter	Rom
15	Paul	Camerino
18	Mary	Olten
25	Joe	Bern



PurchaseID	CustomerID	Date	ProductID
1002	11	5 May 2015	SE4256
1003	18	5 May 2015	EA4516
1004	11	6 May 2015	EA4516
1005	25	6 May 2015	RG3452



ComplaintID	Complaint	Person
36536	Return	George
44363	Failure	Paul
46344	Failure	John

ETL – Harmonisation: Schema integration

Problem	characteristics	Example: data source 1	Example: data source 2	Solution
Synonyms	Attributes with different names have identical meaning	Attribute «employee» contains employee name	Attribute «staff» contains employee name	Choose an attribute name
Homonyms	Same attribute name refers to attributes with different meaning	Attribute «partner» refers to name of customer	Attribute «partner» refers to name of supplier	Choose different attribute names

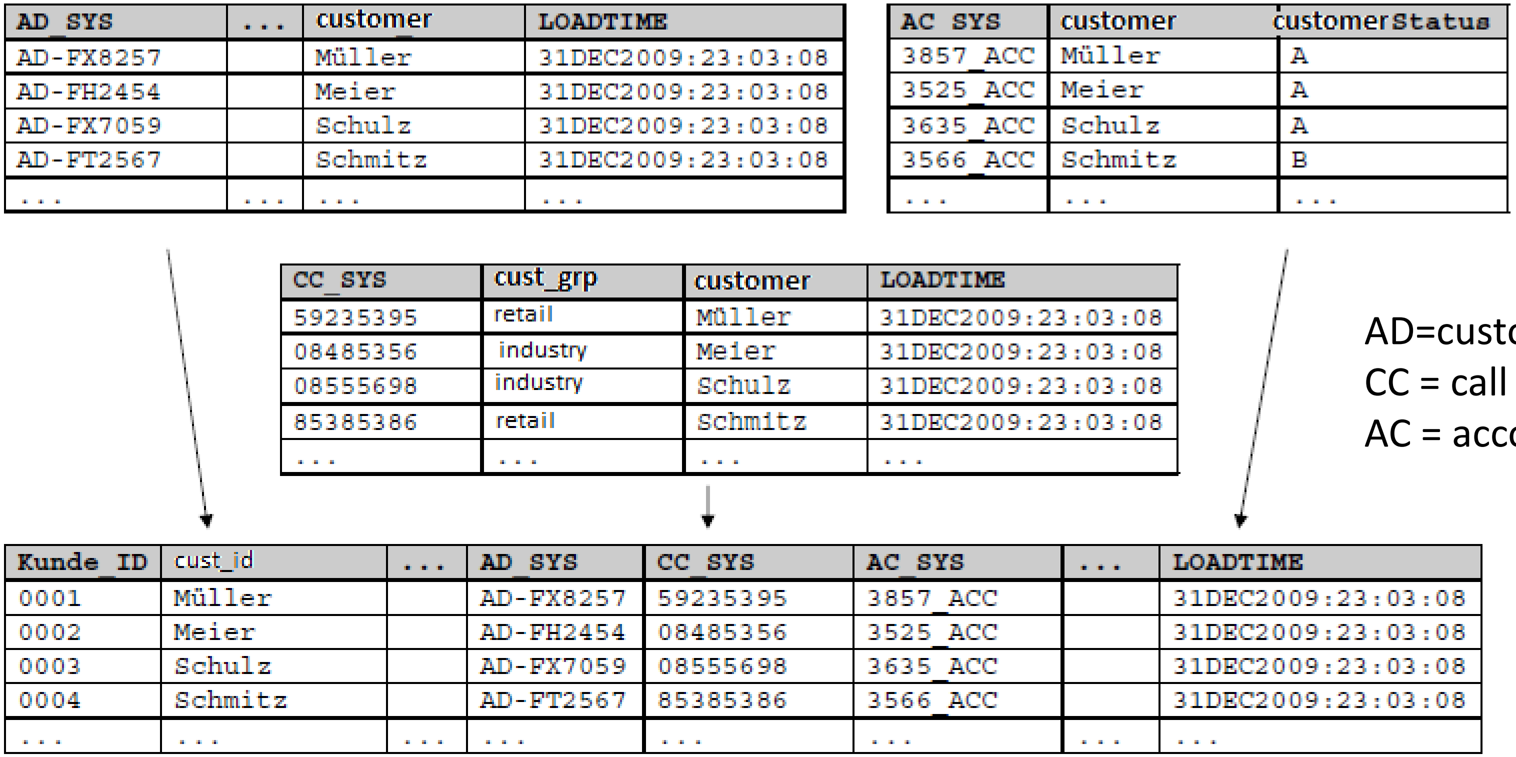
ETL – Harmonisation: data integration (1)

Problem	characteristics	Example: data source 1	Example: data source 2	Solution
Deviating primary keys (synonyms)	Same entity has different id in different operational DBs	Customer «Smith» has id 376_ACC in accounting application	Customer «Smith» has id 7843_CC in call center application	Record linkage: identify identical entities via overlapping attribute values; use mapping table

- How to detect entity identity?

ETL – Harmonisation: data integration (2)

- Mapping tables: allow to map updates in sources to DWH records



AD_SYS	...	customer	LOADTIME
AD-FX8257		Müller	31DEC2009:23:03:08
AD-FH2454		Meier	31DEC2009:23:03:08
AD-FX7059		Schulz	31DEC2009:23:03:08
AD-FT2567		Schmitz	31DEC2009:23:03:08
...	...	...	...

AC_SYS	customer	customerStatus
3857_ACC	Müller	A
3525_ACC	Meier	A
3635_ACC	Schulz	A
3566_ACC	Schmitz	B
...	...	...

CC_SYS	cust_grp	customer	LOADTIME
59235395	retail	Müller	31DEC2009:23:03:08
08485356	industry	Meier	31DEC2009:23:03:08
08555698	industry	Schulz	31DEC2009:23:03:08
85385386	retail	Schmitz	31DEC2009:23:03:08
...	...	...	...

Kunde_ID	cust_id	...	AD_SYS	CC_SYS	AC_SYS	...	LOADTIME
0001	Müller		AD-FX8257	59235395	3857_ACC		31DEC2009:23:03:08
0002	Meier		AD-FH2454	08485356	3525_ACC		31DEC2009:23:03:08
0003	Schulz		AD-FX7059	08555698	3635_ACC		31DEC2009:23:03:08
0004	Schmitz		AD-FT2567	85385386	3566_ACC		31DEC2009:23:03:08
...	...	...	...	...	...	...	...

AD=customer service
CC = call center
AC = accounting

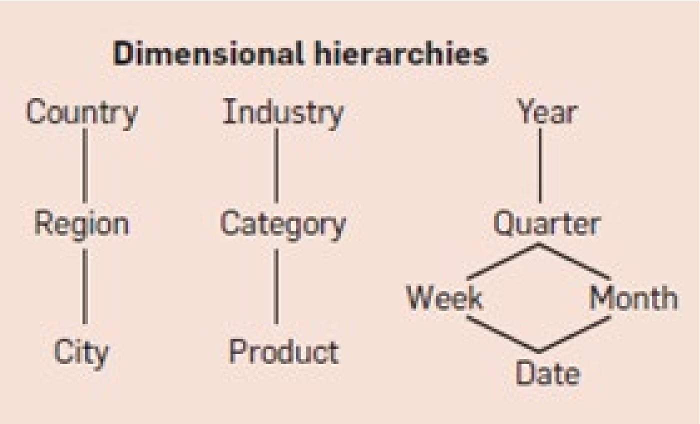
adapted from Kemper et al.

ETL – business harmonisation

- **adjust figures/values**
  - consolidate figures from various databases based on their (business) meaning, e.g. apply rules to map location- or department-specific value deviations
  - convert currencies and units (e.g. inch → cm)
- **adjust granularity**
  - decide for a level of granularity (e.g. monthly or quarterly)
  - harmonise according to period (source systems may have differing granularity, e.g. quarters vs. years)
  - aggregate all values on that level (e.g. sum all records/receipts of one day together)

ETL - Aggregation

- Aggregate data based on dimensional hierarchy
  - usually, aggregates are pre-computed for performance reasons
  - introduces «controlled redundancy»
  - aggregates become invalid when hierarchies and/or source data change…

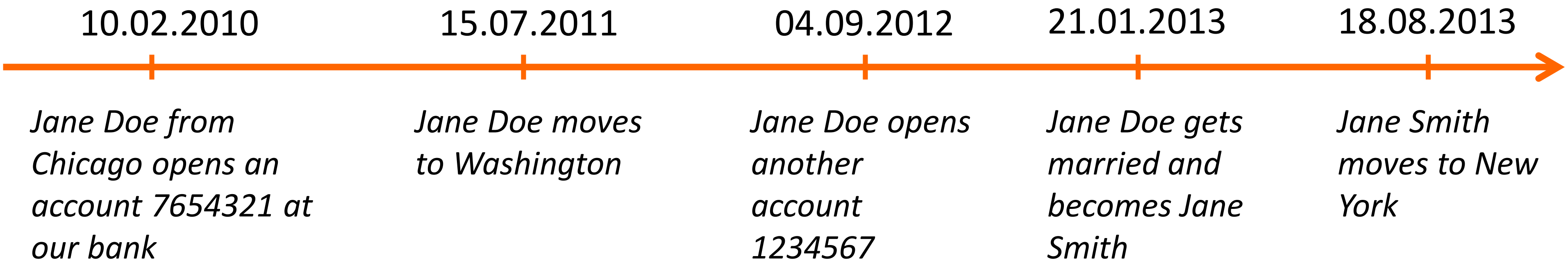


ETL - Enrichment

- add new attributes that are functions of existing data; compute these functions and store the result
  - sums, averages or more complicated computations (e.g. profitability)
  - based on harmonised and/or aggregated data
  - same motivation as aggregation: performance
  - introduces another «controlled redundancy»

Slowly Changing Dimensions

Example: customer dimension change



10.02.2010 — *Jane Doe from Chicago opens an account 7654321 at our bank*

15.07.2011 — *Jane Doe moves to Washington*

04.09.2012 — *Jane Doe opens another account 1234567*

21.01.2013 — *Jane Doe gets married and becomes Jane Smith*

18.08.2013 — *Jane Smith moves to New York*

- who's the owner of the bank account 1234567?
  - as of today: Jane Smith from New York
  - as of 31.12.2012: Jane Doe from Washington
  - as of 31.12.2011: there is no such bank account

Type I: no history

Cust_id	Cust_name	Cust_city	…
1	John Allan	Chicago	…
2	Chris Lee	Boston	…
3	Jane Doe	Chicago	…

Cust_id	Cust_name	Cust_city	…
1	John Allan	Chicago	…
2	Chris Lee	Boston	…
3	Jane Doe	Washington	…

15.07.2011

old value is simple overwritten with new value

Type II: full history

Cust_id	Cust_name	Cust_city	Valid from	Valid to
1	John Allan	Chicago	10.03.2008	
2	Chris Lee	Boston	02.06.2010	
3	Jane Doe	Chicago	10.02.2010	

18.08.2013

Cust_id	Cust_name	Cust_city	Valid from	Valid to
1	John Allan	Chicago	…	
2	Chris Lee	Boston	…	
3	Jane Doe	Chicago	10.02.2010	14.07.2011
3	Jane Doe	Washington	15.07.2011	20.01.2013
3	Jane Smith	Washington	21.01.2013	17.08.2013
3	Jane Smith	New York	18.08.2013	

- every intermediate state is documented, validity range of values is signalled via «valid from», «valid to» attributes
- «valid from» becomes part of primary key

Type III: limited history

Cust _id	Previous Cust_name	Current Cust_name	Effective date cust_name	Previous Cust_city	Current Cust_city	Effective date cust_city
1		John Allan			Chicago	10.03.2008
2		Chris Lee			Boston	02.06.2010
3		Jane Doe			Chicago	10.02.2010

18.08.2013

Cust _id	Previous Cust_name	Current Cust_name	Effective date cust_name	Previous Cust_city	Current Cust_city	Effective date cust_city
1		John Allan			Chicago	10.03.2008
2		Chris Lee			Boston	02.06.2010
3	Jane Doe	Jane Smith	21.01.2013	Washington	New York	18.08.2013

- keeps the n previous values, each in a separate new column (in the example: n=1)
- effective date column(s) show(s) when the change occured

Exercise:

- The company Foobar produces and sells computer hardware. They wish to gain a unified view over their customer base. In the planned data warehouse (DWH), information from the following source systems should be integrated:
  - a call center application (CRM)
  - an ERP that stores orders and invoices (among other things)
  - a campaign management system
- Foobar faces several challenges in the process of building the DWH.
- Assign each of the following examples to the appropriate problem categories

Transformation – Problem 1

- In the CRM, there is a field where call center agents can select the category of the component that is causing problems from a drop-down list. Because of time constraints, agents often fail to fill this in. The task is to try to estimate the component from other fields of a call record, e.g. the description.

□ Filtering (remove duplicates)
□ Filtering (spelling correction)
□ Filtering (missing values)
□ Filtering (detect and handle inconsistencies)

□ Harmonisation (schema integration)
□ Harmonisation (data integration)
□ Business harmonisation
□ Aggregation
□ Enrichment

Transformation – Problem 2

- Customers have the same ids in the CRM and ERP, but different ids in the campaign management application. Task: records referring to the same customer should be identified

□ Filtering (remove duplicates)
□ Filtering (spelling correction)
□ Filtering (missing values)
□ Filtering (detect and handle inconsistencies)

□ Harmonisation (schema integration)
□ Harmonisation (data integration)
□ Business harmonisation
□ Aggregation
□ Enrichment

Transformation – Problem 3

- In some orders in the ERP, the shipping date is earlier than the order date; the task is to spot these records and eliminate them

□ Filtering (remove duplicates)
□ Filtering (spelling correction)
□ Filtering (missing values)
□ Filtering (detect and handle inconsistencies)

□ Harmonisation (schema integration)
□ Harmonisation (data integration)
□ Business harmonisation
□ Aggregation
□ Enrichment

Transformation – Problem 4

- In order to speed up analyses later on, the customer lifetime value of each customer should be pre- computed and stored in the data warehouse

□ Filtering (remove duplicates)
□ Filtering (spelling correction)
□ Filtering (missing values)
□ Filtering (detect and handle inconsistencies)

□ Harmonisation (schema integration)
□ Harmonisation (data integration)
□ Business harmonisation
□ Aggregation
□ Enrichment

Transformation – Problem 5

- Some of the attributes of customers in the CRM are also available in the campaign management system, but have different names, e.g. there is an attribute "status" with possible values "active" and inactive" in the CRM and an attribute "active" with possible values "yes" and "no" in the campaign management system. The task is to define the attributes of the target DWH table where these attributes are brought together

□ Filtering (remove duplicates)
□ Filtering (spelling correction)
□ Filtering (missing values)
□ Filtering (detect and handle inconsistencies)

□ Harmonisation (schema integration)
□ Harmonisation (data integration)
□ Business harmonisation
□ Aggregation
□ Enrichment

Transformation – Problem 6

- Customers can place orders either through a web form or by writing an email. Discounts or price negotiations are not possible when ordering online, therefore, larger orders are often placed via email, and price negotiations may take place. In that case, the responsible sales person enters the ordered items and the total price of the order into the ERP system. The total price may not be equal to the sum of the prices of the purchased items (due to discounts). The task is to define how the customers' purchases should be stored in the target structure of the DWH and at which granularity the information about revenue from these purchases is to be kept

□ Filtering (remove duplicates)
□ Filtering (spelling correction)
□ Filtering (missing values)
□ Filtering (detect and handle inconsistencies)

□ Harmonisation (schema integration)
□ Harmonisation (data integration)
□ Business harmonisation
□ Aggregation
□ Enrichment