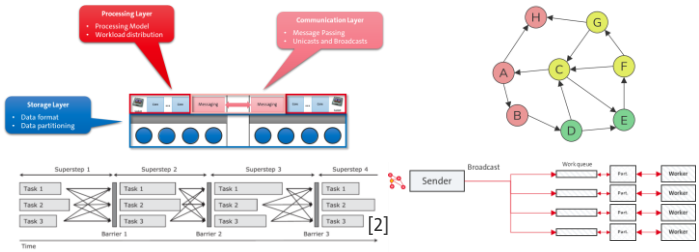
Summary

- Architecture of Database Systems
- Transaction Management
- Modern Database Technology
- Data Warehouses and OLAP
- Data Mining
- Big Data Analytics

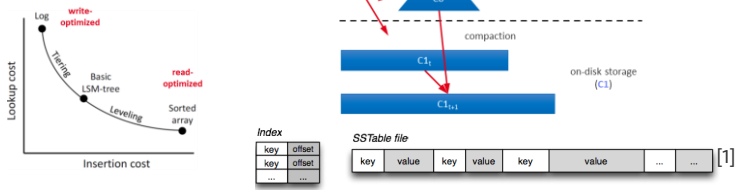
Challenges for Graph Engines



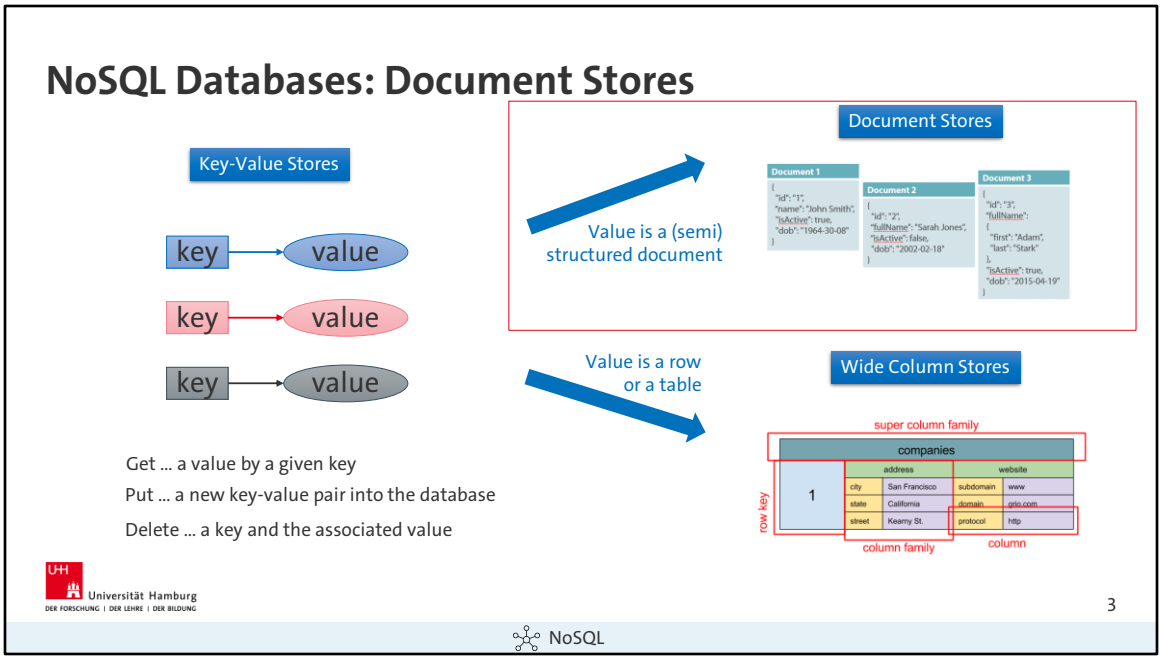
Key-Value Stores

Memcached via telnet

Add a new KV-pair
set AgeAlice 0 120 1 [Press Enter]
26 [Press Enter]
Retrieve a KV-pair
get AgeAlice
Delete a KV-pair
delete AgeAlice



[1] Image: https://www.igvita.com/2012/02/06/sstable-and-log-structured-storage-leveldb/
[2] Robert Ryan McCune, et al.: Thinking Like a Vertex: A Survey of Vertex-Centric Frameworks for Large-Scale Distributed Graph Processing. ACM Comput. Surv. 48(2): 25:1-25:39 (2015),

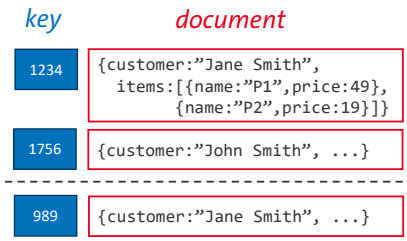


Basic Data Model

- The general notion of a document – words, phrases, sentences, paragraphs, sections, subsections, footnotes, etc.
- Flexible schema: subcomponent structure may be nested, and vary from document-to-document
- Essentially, they support the embedding of documents and arrays within other documents and arrays
- Document structures do not have to be predefined, so are schema-free (XML, mostly JSON)

Document Stores

- Application-oriented management of structured, semi-structured, and unstructured information
- Collections of (key, document)-pairs



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

4

NoSQL

Motivation

- Application-oriented management of structured, semi-structured, and unstructured information
- Scalability via parallelization on commodity HW (cloud computing)

System Architecture

- Collections of (key, document)
- Scalability via sharding (horizontal partitioning)
- Custom SQL-like or functional query languages

Example Systems

- MongoDB (C++, 2007, CP) → RethinkDB, Espresso, Amazon DocumentDB (Jan 2019)
- CouchDB (Erlang, 2005, AP) → CouchBase

Recap: JSON (JavaScript Object Notation)

JSON Data Model

- Data exchange format for semi-structured data
- Not as verbose as XML (especially for arrays)
- Popular format

```
{“students:”[
  {“id”: 1, “courses”:[
    {“id“:“INF.01017UF”, “name“:“DM”},
    {“id“:“706.550”, “name“:“AMLS”}]},
 {“id”: 5, “courses”:[
    {“id“:“706.520”, “name“:“DIA”}]},
]}
```

Query Languages

- Most common: libraries for tree traversal and data extraction
- JSONig: XQuery-like query language
- JSONPath: XPath-like query language

JSONiq Example:

```
declare option jsoniq-version “…”;
for $x in collection(“students”)
  where $x.id lt 10
  let $c := count($x.courses)
  return {“sid”:$x.id, “count”:$c}
```

[http://www.jsoniq.org/docs/JSONiq/html-single/index.html]

Example MongoDB

[**Credit:** https://api.mongodb.com/python/current]

Creating a Collection

```
import pymongo as m
conn = m.MongoClient("mongodb://localhost:123/")
db = conn["dbs19"]      # database dbs19
cust = db["customers"]  # collection customers
```

Inserting into a Collection

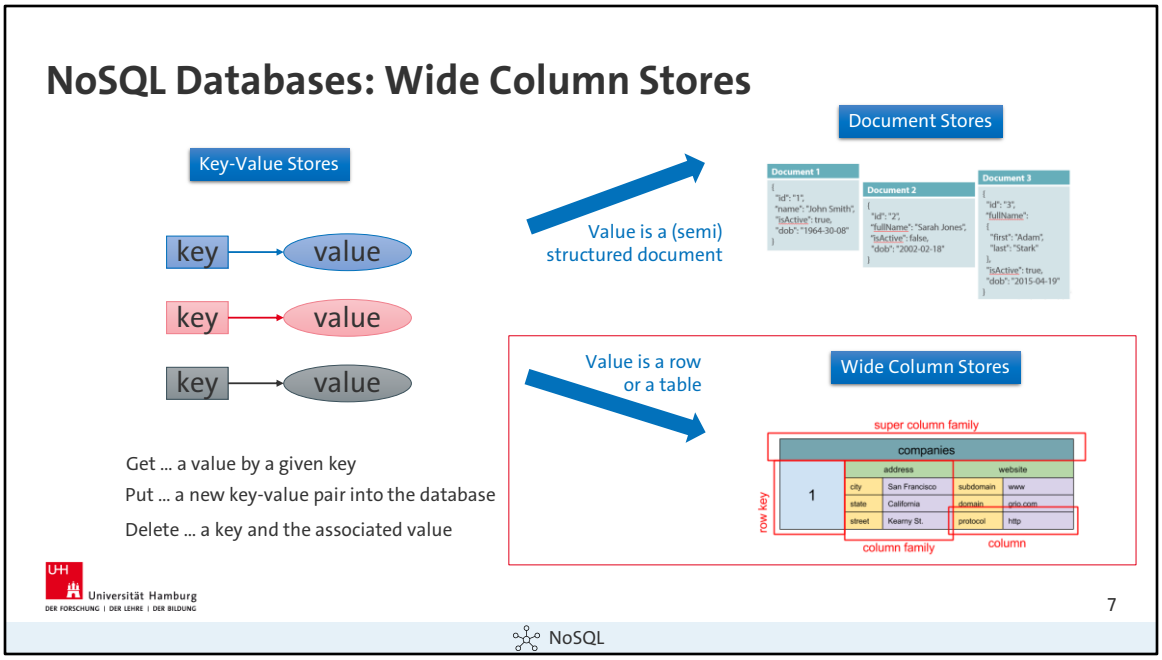
```
mdict = {
  "name": "Jane Smith",
  "address": "Inffeldgasse 13, Graz"
}
id = cust.insert_one(mdict).inserted_id
# ids = cust.insert_many(mlist).inserted_ids
```

Querying a Collection

```
print(cust.find_one({"_id": id}))

ret = cust.find({"name": "Jane Smith"})
for x in ret:
  print(x)
```

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



- Wide Column Stores are sometimes called “Extensible Record Stores”
- Column Store != Wide Column Store

Basic Data Model

- Database is a collection of key/value pairs
- Key consists of 3 parts: a row key, a column key, and a time-stamp (i.e., the version)
- Flexible schema: the set of columns is not fixed, and may differ from row-to-row

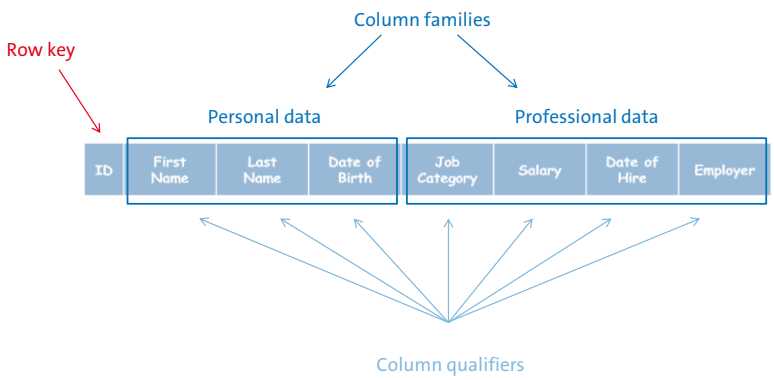
Example Systems

- Google Bigtable
- HBase

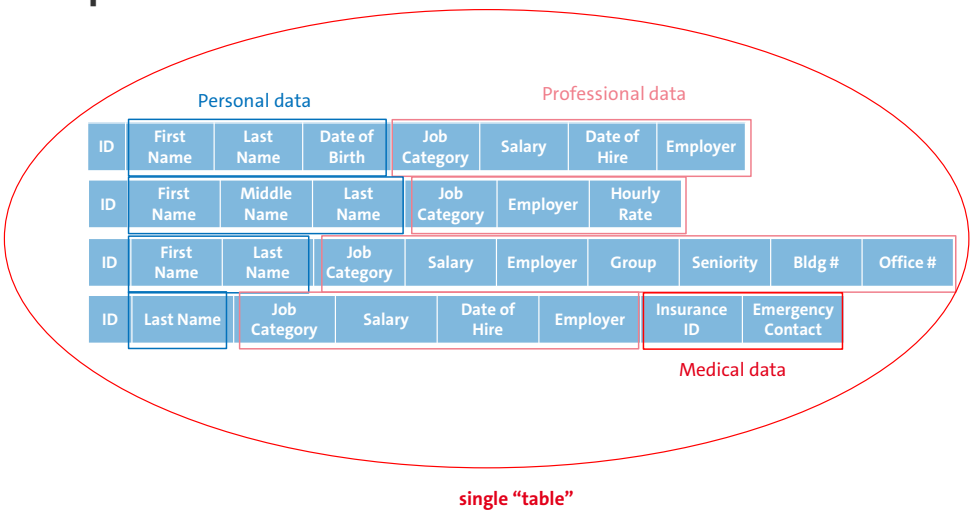
Further Reading

Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." *ACM Transactions on Computer Systems (TOCS)* 26.2 (2008): 1-26.

Example: Wide Column Data Model

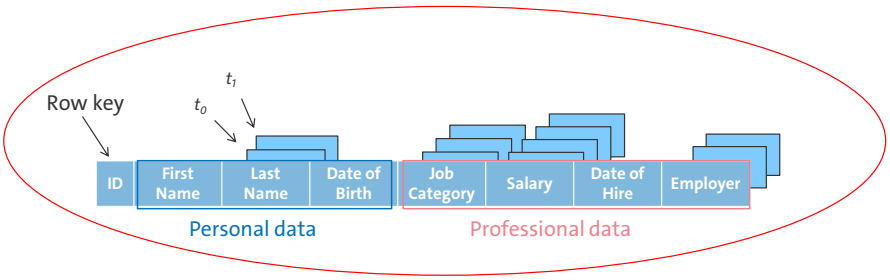


Example: Wide Column Data Model



single "table"

Example: Wide Column Data Model



One "row"

One "row" in a wide-column NoSQL database table
=
Many rows in several relations/tables in a relational database

Overview NoSQL Systems

Key-Value

LEVELDB
MEMCACHED
redis

Document

CouchDB
mongoDB

Wide Table

APACHE HBASE
Cassandra

Graphs

APACHE GIRAPH
neo4j
ArangoDB

Time Series, Spatial, Arrays...

influxdb
Prometheus
druid
SciDB
TIMESCALE

11

Further Reading

Ronald Barber et al: Evolving Databases for New-Gen Big Data Applications. CIDR 2017

Course Outline

- Architecture of Database Systems
- Transaction Management
- Modern Database Technology
- Data Warehouses and OLAP
- Data Mining
- Big Data Analytics

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

12

Motivation for a Data Warehouse

Preparation

- Multitude of different data sources
- Data cleansing / cleaning → historic reason for DWH

Adaptation/Standardization/Interpretability

- Central defined key performance indicators (KPI Management) and dimensions
- One single truth, no different interpretations

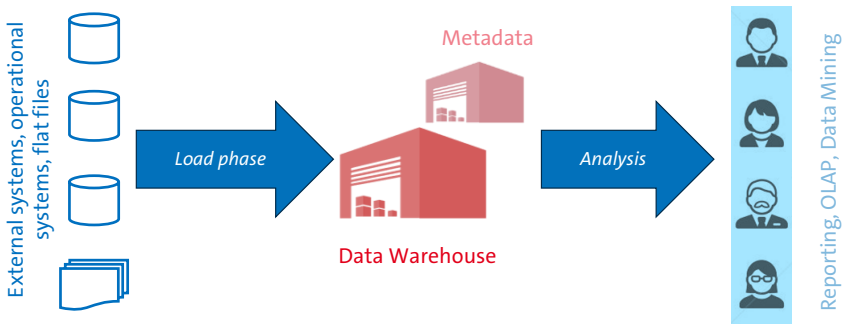
Allows new business processes

- Statistical methods (e.g. correlation analysis)
- Examples: customer segmentation, -evaluation

More

- Organization: data from multiple business units
- Technical: different query models, usage patterns, data volumes
- Consolidation: heterogeneity (schema, data quality, “garbage in, garbage out!”)

Data-Warehouse concept



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

13

Data Warehouses and OLAP

Goal: (analytic) access to consistent data

- Integration of external/operational data sources in a centralized database
- Extraction and multi-stage loading process
- Integrated (and historicized) data form a base for analytic tools

Descriptive Statistics

Processing step	Actions	Data
Data recording	Creation of raw data calibration / checking optional: anonymization	Raw data „Microcosmos“
Data preparation	Structural adaptation Validation (filtering, quality checking) Statistical corrections (imputation, estimation, „missing values“) Optional: anonymization, pre-aggregation	processed raw data
Data analysis	Application-specific aggregation Representation / interpretation	Enriched data/aggregates, „Macrocosmos“

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

14

Data Warehouses and OLAP

Explorative statistics

- Find “undiscovered” structures and connections to generate new hypotheses
- Based on samples

Mathematical statistics

- Also called “statistical inference” or “inductive statistics”, in German also “schließende Statistik”
- Based on samples
- Uses stochastic models → provides probability of error (in contrast to descriptive statistics)

DWH Definition

A data warehouse is a central database which is optimized for analysis purposes and contains data of several, usually heterogeneous data sources.

Erhard Rahm

Physical database as an integrated view on (arbitrary) data with a focus on the evaluation aspect, where data is often but not necessarily historized."

GI

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

15

Data Warehouses and OLAP

Some general approaches

"Data warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data in support of the management's decision-making process." (Bill Inmon)

"Data Warehouse is an environment, not a product." (Berson/ Smith)

Characteristics of DWHs

Analysis-oriented organization of data

Domain-oriented, models a specific application goal

Integration of data from different source systems

Integrated database, integration on structural and data level of multiple databases

No user updates

(almost) **no updates and deletes** (technical updates / deletes only, quality assurance)

Historic data

Data is kept over a long period of time

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

16

Data Warehouses and OLAP

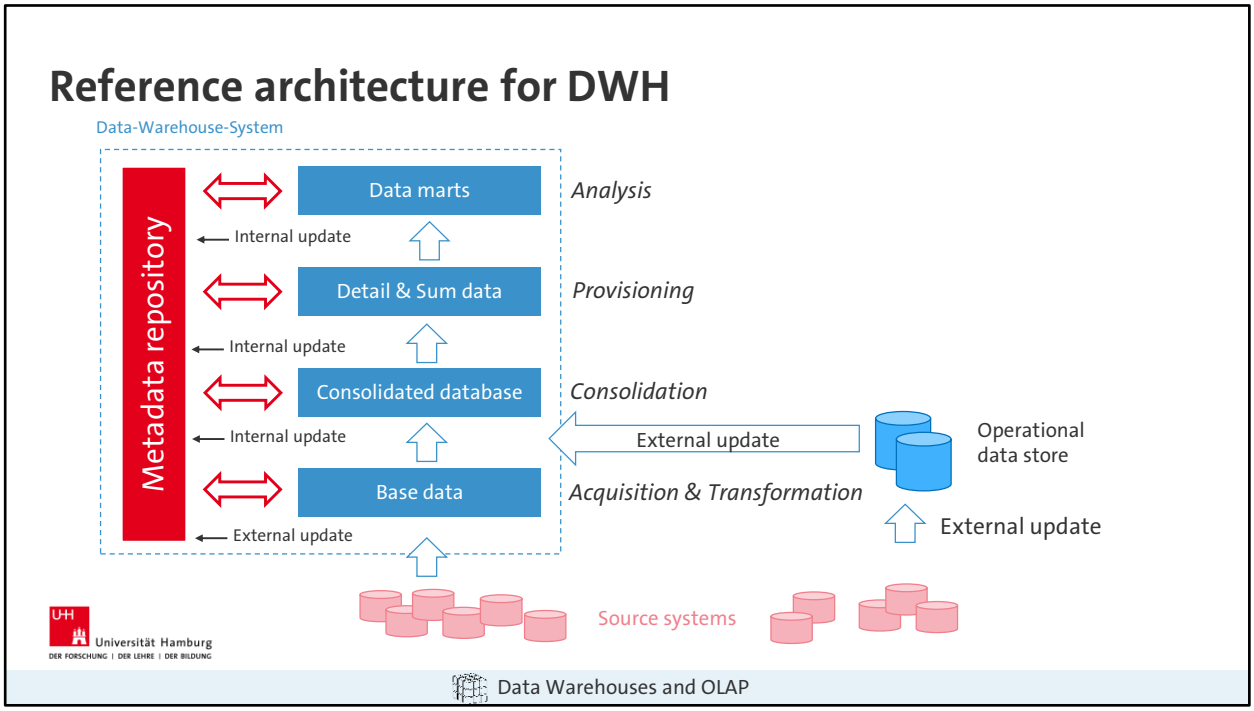
- Operations in operational environment: Insert, Delete, Update, Select
- Operations in a data warehouse: Insert the initial and additional loading of data by (batch) processes, Select the access of data
- Non-volatile/stable database, loaded data is not deleted or modified, read-only access

<u>Non-volatile Data</u>

- What happens in the OLTP system if the customer cancels his booking?
    - Delete operation in OLTP
    - Seat gets available again and can be sold to another passenger
- What happens in the DWH?
    - Insert operation in DWH with, e.g., a flag indicating that the customer cancelled/deleted his booking
    - Business can make analysis about cancelled booking: why might the customer have cancelled? How to prevent the customer or other customers to cancel next time?

Tales from the reality of data science

- “Compulsive data hoarders” exist in many (but not all) projects, so data is collected before it is defined what this data will be used for or if it will be used at all → Exponential data growth
- Especially in natural science, it’s not always clear which parts of the data might be useful at a later point in time → Everything is collected → Data growth is only restricted by the ratio of failed experiments/failed hardware



Further Reading

Wolfgang Lehner: „Datenbanktechnologie für Data-Warehouse-Systeme: Konzepte und Methoden“, dpunkt-Verlag, 2003

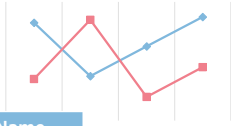
Metadata examples

- Description of the Data Warehouse System
- Names, definitions, structure, and content of a DWH
- Identification of data sources
- Integration and transformation rules for filling the DWH
- Integration and transformation rules for end-user rules
- Operational information like updates, versions
- Usage and performance of the Data Warehouse (Monitoring)
- Security, access rights

Example DWH architecture

Data analytics

ZID	MID	Salary	ZID	MID	Salary
1	1	3.300€	1	2	6.176€
2	1	3.500€	2	2	5.176€



Data provisioning

ZID	Time	ZID	MID	Salary	MID	Name
1	T1	1	1	3.300€	1	Max Mueller
2	T2	1	2	6.176€	2	Tim Mueller
		2	1	3.500€		
		2	2	5.176€		

Data consolidation

Zeit	Name	Salary
T1	Max Mueller	3.300€
T1	Tim Mueller	6.176€
T2	Max Mueller	3.500€
T2	Tim Mueller	5.176€

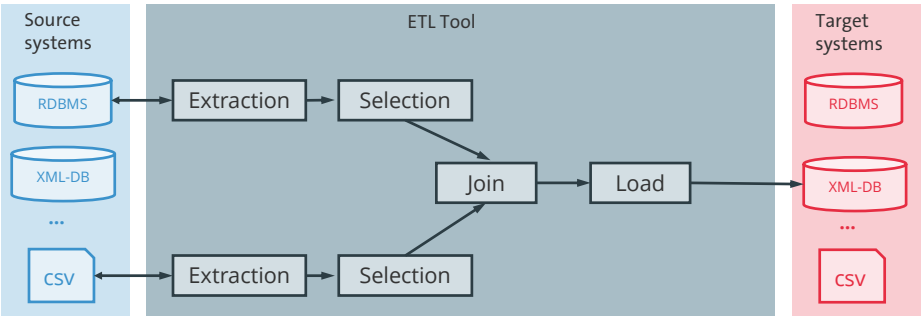
Data transformation

Sources

Name		Salary	
Name	Gehalt	Name	Salary
Max Müller	3.500€	Tim Mueller	7,400$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Data acquisition and transformation



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

19

Data Warehouses and OLAP

Goal
- Provisioning of data for the consolidated database
- Efficiency vs. Actuality

What are data sources?
- Heterogeneous systems or files
- Local schemata and semantics

Describing attributes of data sources
- Utilization of data for the Data Warehouse
- Origin (internal or external data)
- Cooperation (active sources, snapshot-sources, ...)
- Availability of source data (legal, social, organizational, technical)
- Cost of acquisition
- Quality (consistency, correctness, completeness, accuracy,...)

Staging Area
- “Landing Zone” for data coming into a DWH
- Temporary memory for integrating extracted data after an external update

Schema integration

- Overcoming semantic/structural heterogeneity
- Integration of different local schemata /data models into one global schema
- Decoupling of transformation from the source systems and the consolidated database

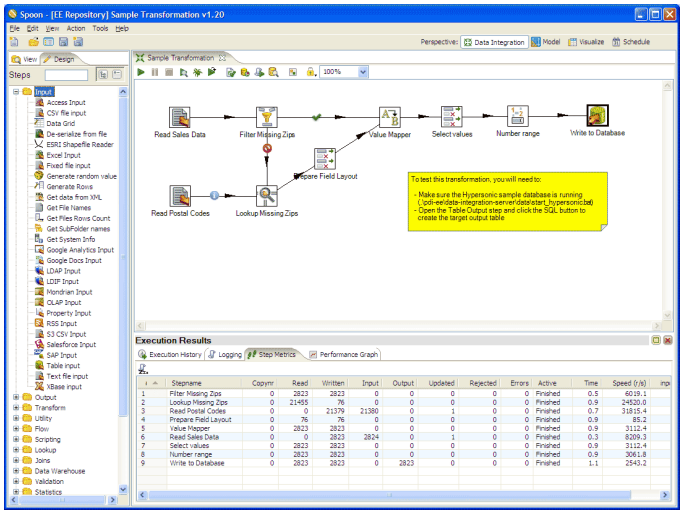
Data integration

- Adaptation of data formats etc.
- Correction of different spellings (abbreviations, etc.)

ETL-components

- Extraction-, Transformation- and Load component

Example ETL Tools



https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho/download-pentaho.html

Commercial Systems

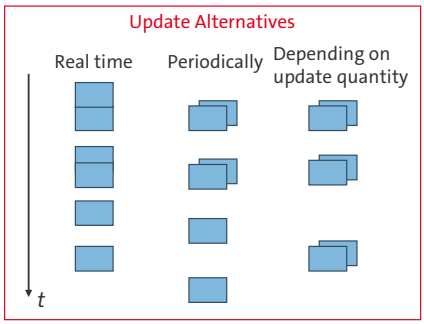
- Informatica PowerCenter
- Cognos Decision-Stream
- Oracle DW Builder
- IBM InfoSphere DataStage
- IBM DB2 Warehouse Enterprise Edition
- AB Initio

Open Source

- Pentaho Data Integration (a.k.a Kettle)
- Talend ETL Integration Suite
- Clover ETL Data Integration
- JasperETL Open Source

Data Consolidation

- Integrated database from cleaned data
- Not specifically modeled, **no specific optimizations**



...for schema (classical logical and physical database design)

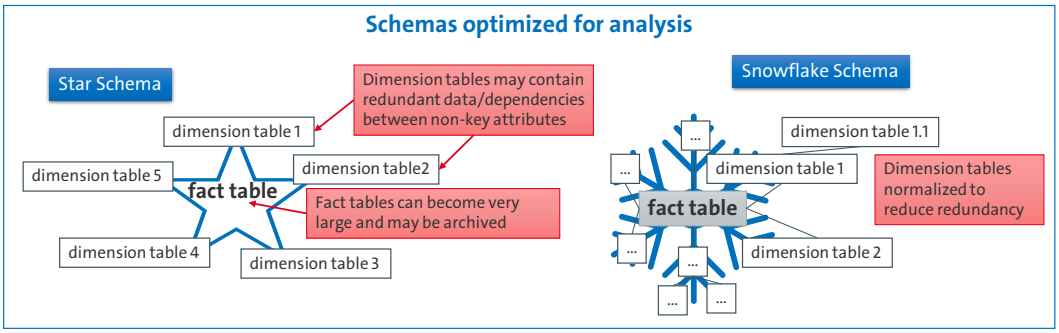
...for physical database design (index structures, etc.)

Function

- Organization-spanning and application-independent data storage
- Collection, integration and distribution
- Analytic functionality for operative use

Data Provisioning and Analysis

- Dispositive database derived from consolidated data
- Mostly historic, maybe detailed data, mostly complete



22

Data Warehouses and OLAP

More optimization for analysis

- Logical access paths
- Partitioning
- Pre-calculation of summed data (e.g. materialized views)
- Physical access paths: specific index structures (e.g. bitmap index)

Data analysis

- Data-Mart databases derived from dispositive database
- Specific extracts for a specific class of applications
- Mostly proprietary formats on the physical level (e.g. MOLAP-systems)

Survey

https://evasys-online.uni-hamburg.de/evasys/online.php?pswd=J3Q6H

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG