Chapter 3:
Data Provisioning

1. Introduction
2. Data extraction
3. From transactional data towards analytical data
4. Schema and data integration

1. Introductrion

It‘s all about the data.[...]But data doesn’t come to you...”
− Data collection,extraction,and integration is often the most complex and expensive tasks in a BI project
− According to Bernstein and Haas2
- “information integration is thought to consume about 40% of their budget”
- “the market for data integration and access software [...] was about $2.5 billion in 2007 and is expected to grow to $3.8 billion in 2012“

1. Introductrion

In addition: more and more data is available

- According to Chauduri et al.3, we face „*very large amounts of data arising from sources such as customer transactions in banking, retail as well as in e-businesses, RFID tags for inventory tracking, email, query logs for Web sites, blogs, and product reviews*”
- On top, “*Real-world Data is Dirty*” according to Hernandes and Stolfo4

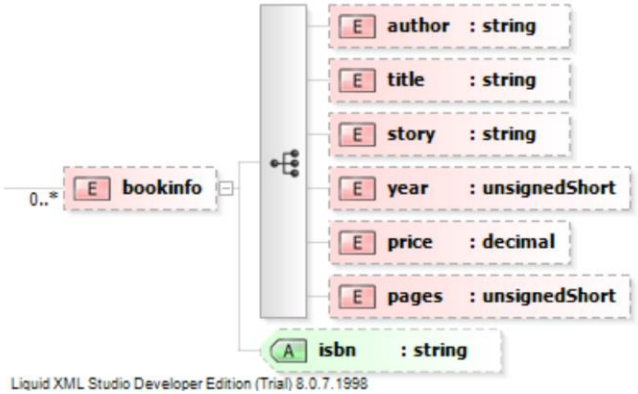
therefore data quality is of utmost importance

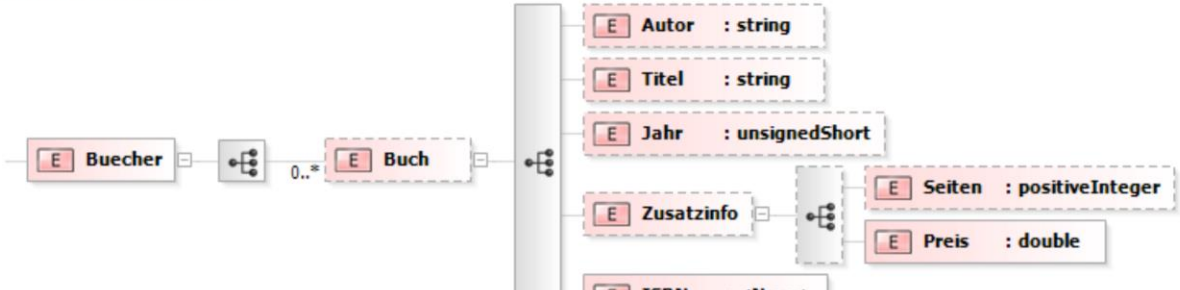
- Crucial: Keep an eye on your analysis goals!
- In summary, we have to

collect/select
extract
clean, and
integrate data

1. Introductrion

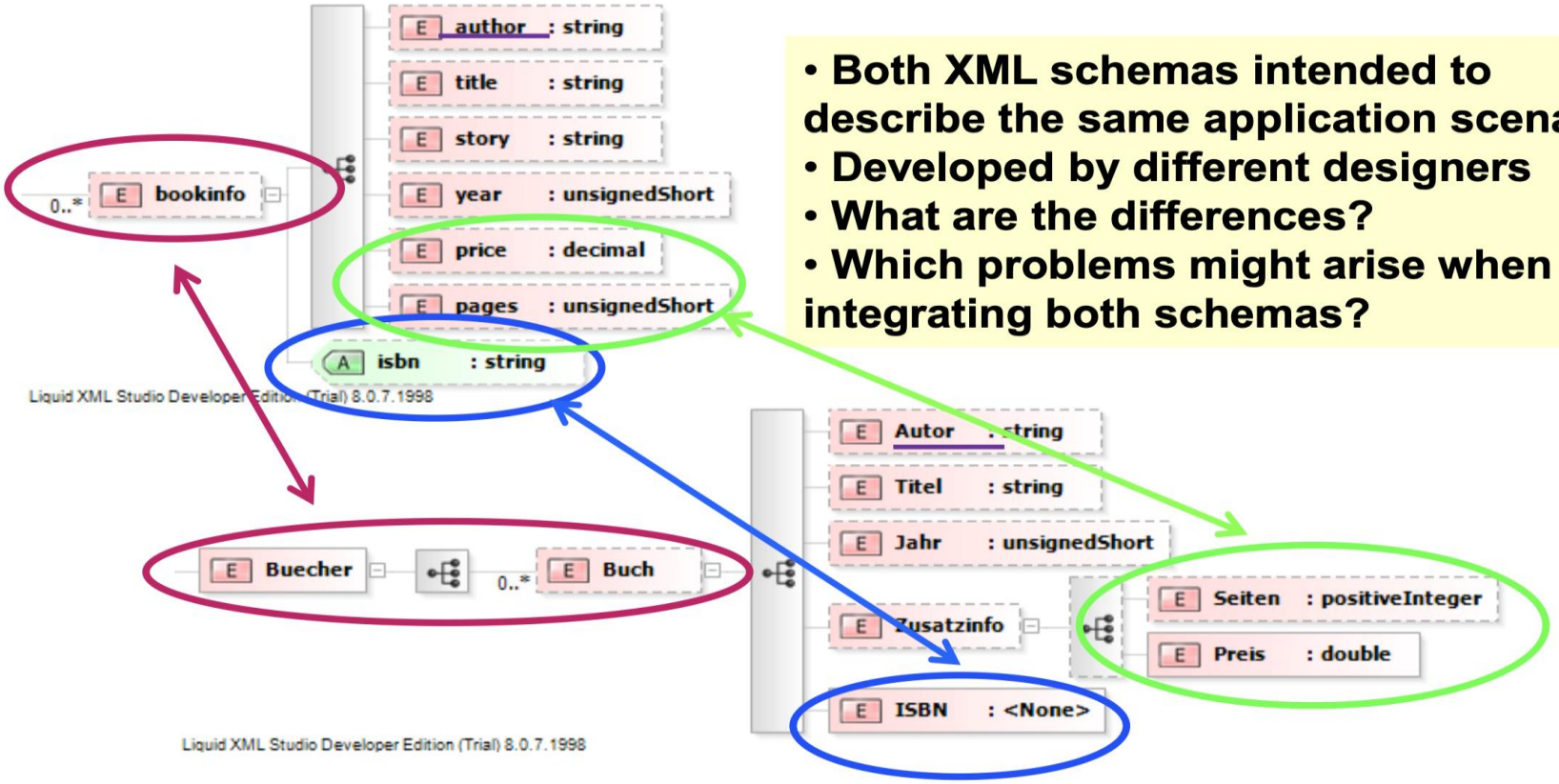
- Example problem:integration at schema level





- **Both XML schemas intended to describe the same application scenario**
- **Developed by different designers**
- **What are the differences?**
- **Which problems might arise when integrating both schemas?**

1. Introductrion



- **Both XML schemas intended to describe the same application scenario**
- **Developed by different designers**
- **What are the differences?**
- **Which problems might arise when integrating both schemas?**

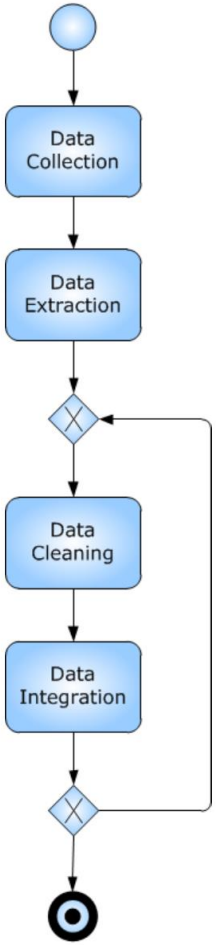
Step 4: Definition of Integration and Cleaning Strategy

Step 3: Analysis of Data (Quality)

Step 2: Definition of Analysis Format (Model)

Step 1: Definition of Analysis Goals

Data Sources



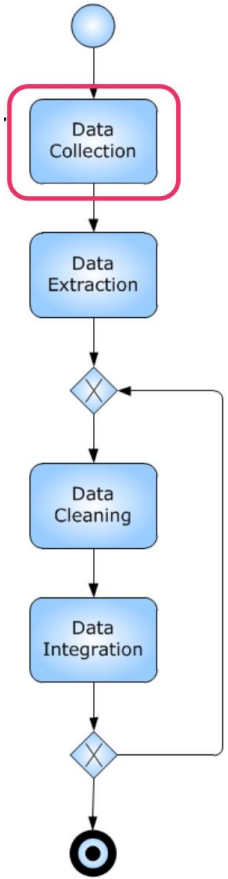
1. Introductrion

This chaper aims at conveying approaches, techniques, and tools to build an integrated data basis for an BI project, in particular:

- Understanding challenges in obtaining and integrating data
- Learning basic techniques of data extraction
- Understanding challenges and learn techniques for improving data quality
- Getting to know different data integration formats
- Understanding how to determine a data integration strategy
- Understanding challenges and learn techniques for data integration in different target formats
- Getting to know use cases from different domains

2. Data extraction

- Remember:„*It‘s all about the data.[...]But data doesn’t come to you*...”
- – In practice different situations
- – Data sources are already existing(and accessible) assumed in literature, practically not always the case →
- – Nonetheless, the relevant sources have to be selected
- – Necessary data is collected „on-demand“ (or in the right
- format)
- – Conclusion1:**Datacollection** is an active task



2. Data extraction

Conclusion 1: **Data collection** is an active task

- Identification of relevant data sources
- Clarification of issues such as data access (particularly, if external data sources are to be accessed)
- **Use Case 1: Patient treatment processes**
- EBMC2 project[5]: co-funded by University of Vienna and Medical University of Vienna
  - » Formalizing medical guidelines for skin cancer treatment
  - » Mining and analysis of real-world treatment processes
  - » In particular regarding their compliance with the guidelines
  - » Selected Key Performance Indicators:
    - Survival time
    - Health status of a specific group of persons
    - Cost effectiveness of certain health policies

2. Data extraction

Balance between:

• What data sources do we need (to fulfill a certain analysis goal) and

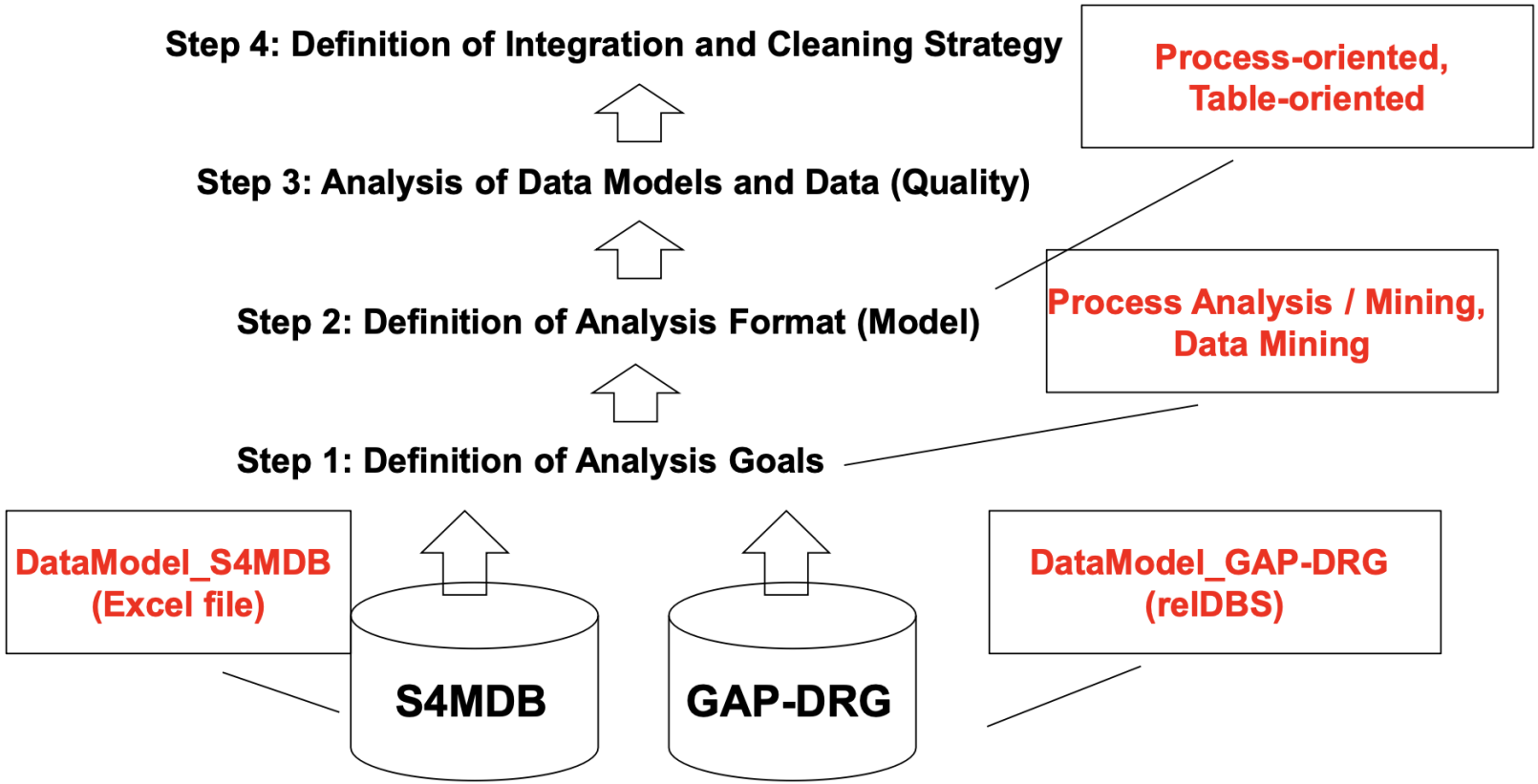
• Which data sources are actually available and accessible (privacy, data ownership, data access costs, etc.)

− Available data sources:

• detailed data collection of clinical Cutaneous Melanoma (CM) stage IV protocols (Stage IV Melanoma Database, S4MDB, for short)

• administrative data of the Main Association of Austrian Social Security Institutions comprising a billing-oriented view of medical patient treatments (GAP-DRG)

2. Data extraction



2. Data extraction

Patient	Id	GivenName	Surname	BirthDate

Treatment	Id	Code	Label

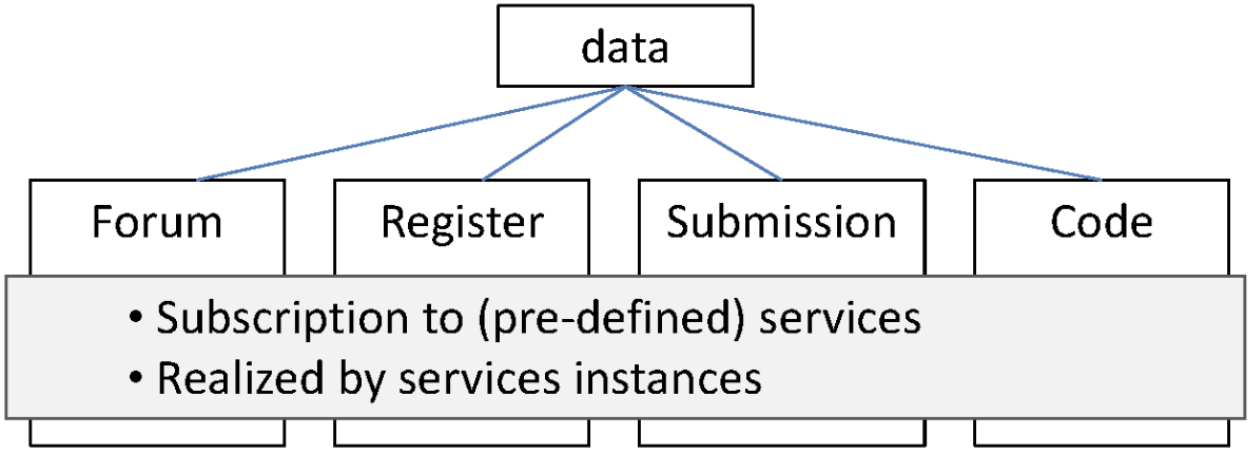
HospitalStay	Id	PatientId	Admission	Discharge

StayTreatment	Id	TreatId	StayId	made

S4MDB

2. Data extraction

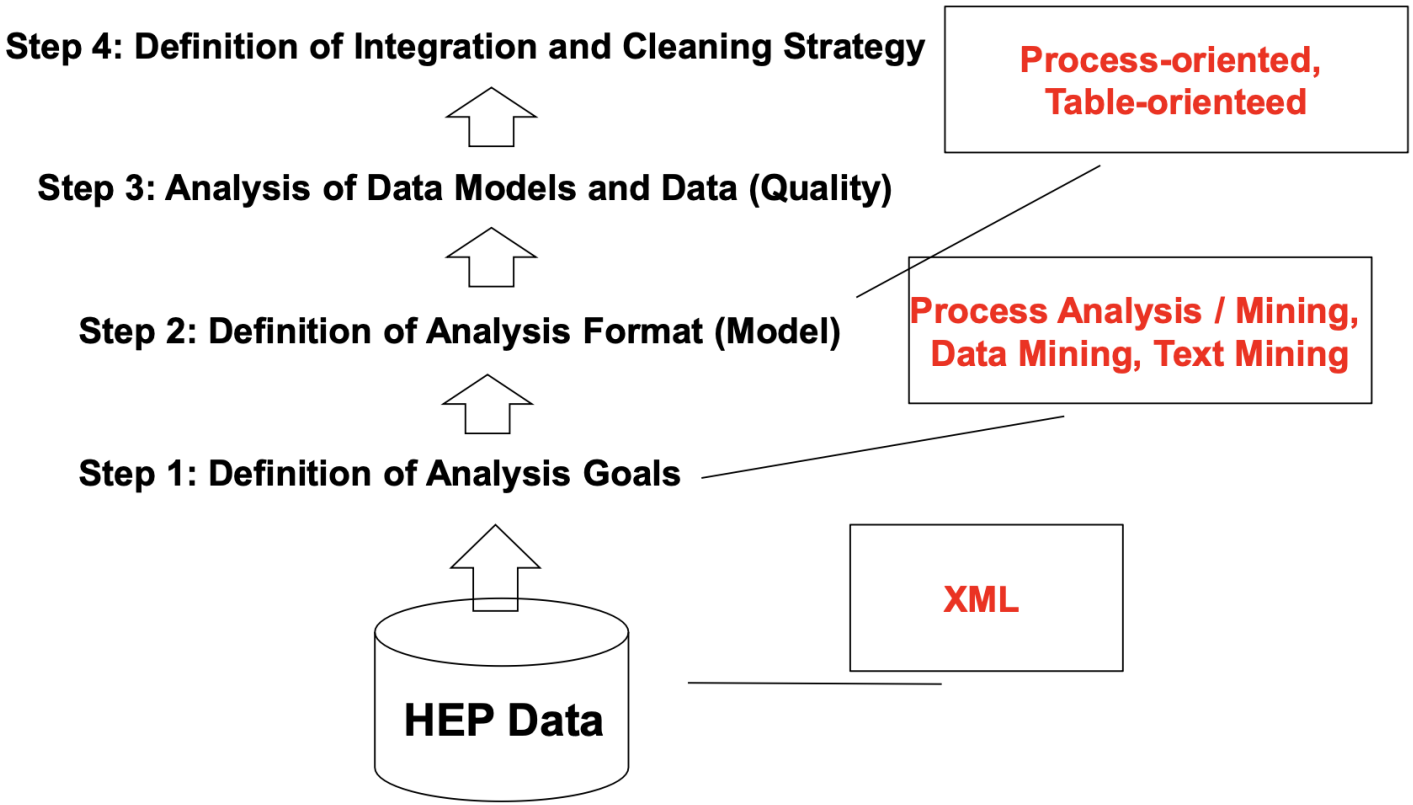
- Use Case 2:Higher-Education Data(HEP)
  - Data source for practical project in Summer
  - Collected from service-oriented learning platform CEWebs



2. Data extraction

- Main analysis questions:
- Analysis of learning processes
- Mining of reference processes
- – Selected key performance indicators:
- Success of learning techniques (e.g., forum)
- Flexibility degree (i.e., analyzing deviations from reference process)

2. Data extraction

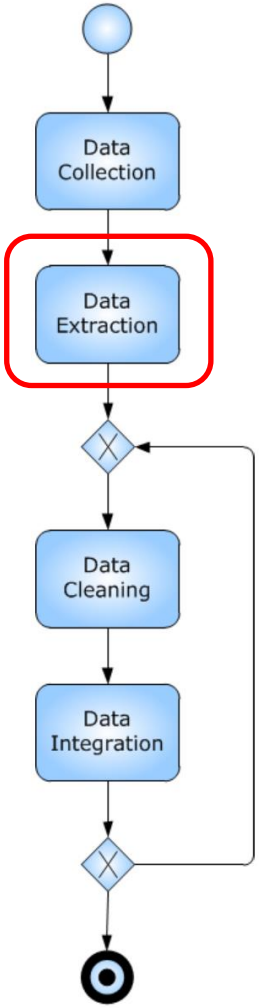


2. Data extraction

- **Further Use Cases, taken from Business Process Intelligence Challenge**
- BPIC 2014: IT-Management:
- Rabobank Group ICT
  Implementation of frequent software releases managed by ITIL processes
- Analysis of underlying change processes to
  predict the workload faced by Service Desks and IT Operations
- – BPIC 2015: Municipalities (NL) - Building Permits
- Collection of building permit application data by several municipalities
- Understand the processes and roles of the participants, and differences in the execution between municipalities
- – BPIC 2016: Customer Contacts
- Employee Insurance Agency (NL)
- Focus on Customers' utilization of various communication channels
- Analysis of the customer behavior

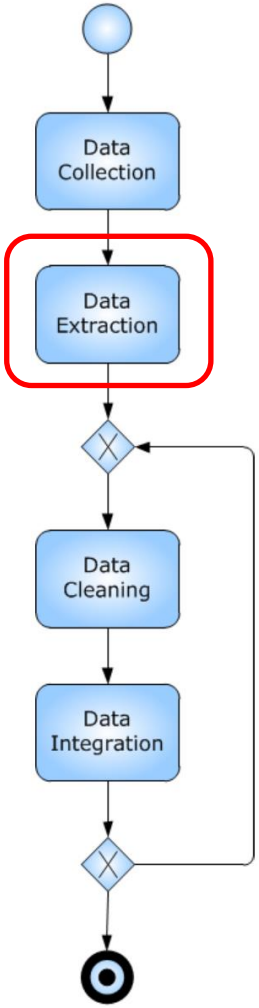
2. Data extraction

- After selecting and/or collecting data sources,data has to be extracted
- Data extraction is a rather technical question:
- Classically: ETL(Extraction–Transformation–Load)
- Access to heterogeneous data sources
- Depends on the type of data source
- Important: do we need the the entire data (or fragments) OR do we need a data update (delta file)?
  - Example (relational) databases: offer access by query language (SQL), but also by logging
  - Example legacy systems: do not offer any support -> many approaches for determining snapshot deltas, e.g., by Window algorithm



2. Data extraction

- Commercial Tools:
- SQL Server Integration Services (included in Microsoft SQL Server product line)
- Oracle Data Integrator
- SAP BusinessObjects Data Integrator
- SAS Data Integration Server
- Open Source / Dual-licensed • Pentaho
- • Talend Open Studio



2. Data extraction

» New Trend: Managing big data

- Computational sciences
- Cloud computing
- Data from social networks
- Sensors

2. Data extraction

According to Beyer challenges are

- Data volume:
  - Data becomes „too big“ for (relational) databases -> Big Tables, NoSQL
  - “Too much volume is a storage issue, but too much data is also a massive analysis issue.” -> MapReduce, BigQuery
- Data velocity:
  - Data extraction during runtime
  - Continuous data streams (e.g., produced by sensors)
- Data Variety:
  - Structured versus unstructured data
  - Cross-sectional vs. event-based data
  - Text, images, videos

2. Data extraction

- *Data volume*
  - NoSQL databases,not based on tables as basic data structures, instead:
    - Document-stores (data variety)
    - Graph databases
    - Key-Value storage systems
- Commercial solutions:
  - Google's Big Table: https://cloud.google.com/bigtable/
  - Amazon's Dynamo: https://aws.amazon.com/de/dynamodb/
  - Facebook's Cassandra
- Open Source solutions:
  - Apache Hadoop
  - Key-Value storage systems

2. Data extraction

» Graph databases

– Before RDBMS:CODASYL and IMS databases(still runningin many enterprises!)

– The data is represented as graph structure

– Queries navigate on the graph structure

– In principle well suited for handling large data sets: WHY?

2. Data extraction

− Example sones GraphDB (sones.de)

• Combining object-oriented aspected and graph

database

• Basic structure: graph G:=(V, E) with V set of vertices and E set of edges

• Definintion in Graphical Query Language (GQL): CREATE VERTEX TYPE Person

ATTRIBUTES (SET<Person> Debitors, SET<Person> Buddies, String name)

INCOMINGEDGES (Person.Debitors ows, Person.Buddies friend Of)

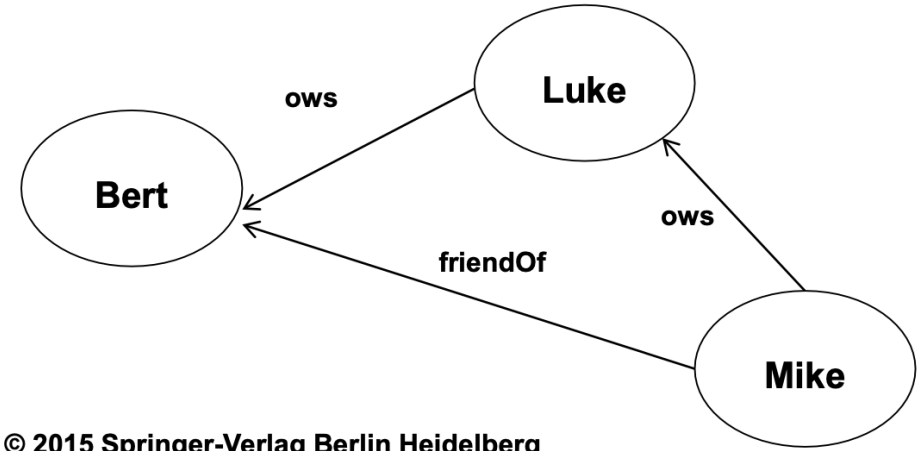
Definition of the vertex types

Set definition → object-orientation

Definition of the edges

2. Data extraction

```
INSERT INTO Person Values (name='Bert')
INSERT INTO Person VALUES (name = "Luke", Debitors =
SETOF(name = "Bert"))
INSERT INTO Person VALUES (name = "Mike", Debitors =
SETOF(name = "Luke"), Buddies = SETOF(name = "Bert"))
```



2. Data extraction

» Graph databases

– Before RDBMS:CODASYL and IMS databases(still runningin many enterprises!)

– The data is represented as graph structure

– Queries navigate on the graph structure

– In principle well suited for handling large data sets: WHY?

2. Data extraction

Queries:

FROM Person

SELECT name, Debitors, Buddies

Selection of Result Mike


```
      "Properties": {
        "name": "Mike"
      }
    },
    {
      "Edges": [
        {
          "HyperEdgeView": {
            "Debitors": [
              {
                "SingleEdge": [
                  {
                    "Properties": []
                  },
                  {
                    "TargetVertex": [
                      {
                        "Properties": {
                          "VertexTypeID": "-9223372036854775782",
                          "VertexID": "-9223372036854775807"
                        }
                      },
                      {
                        "Edges": []
                      }
                    ]
                  }
                ]
              }
            ]
          }
        },
        {
          "HyperEdgeView": {
            "Buddies": [
              {
                "SingleEdge": [
                  {
                    "Properties": []
                  },
                  {
                    "TargetVertex": [
```

2. Data extraction

- Key-Value storage systems
- According to Agrawaletal.,they are
  - adopted by various enterprises.
  - Data analysis: MapReduce paradigm
  - open-source implementation Hadoop
  - widespread adoption in industry and academia
  - Solutions to improve Hadoop systems’ usability and performance

2. Data extraction

Data Variety

– Document-stores

Ready for storing unstructured data

1st possibility: XML extensions on relational DBMS (SQLXML standard)

- Example DB2 Express
- New type XML
- Can be queried using Xpath

By contrast: storing documents as CLOB, however, limited query functionalities (retrieval)

2nd possibility: XML databases Example BaseX (http://basex.org/)

Stores XML files containing structured and unstructured, i.e., document-oriented content

2. Data extraction

» Summary:

– Data variety / data heterogenity is an old and new problem

– Data extraction is a technical question, however, thoughts on data quality and later integration strategy are crucial

– Myriad of tools offer support

– However, definition and implementation of data cleaning and integration strategies (including mapping and definition of target formats) is manual job

– Tools support the definition, documentation of the process as well as support maintenance in case of changes

2. Data extraction

- Summary:
- − New challenges mainly in data velocity, i.e., just-in-time data extraction becomes necessary
- − Big data volume has led to looking for NoSQL databases such as Graph databases, Key/Value stores, document databases
- − By contrast: extensions of RDBMS, Big Tables, etc.
- − After discussion of data extraction techniques, crucial to
- discuss integration formats and data quality issues