2. November 2010

(Advanced) Cloud Computing

Teamproject & Project

Organizer: Prof. Dr. Georg Lausen

Advisors: Alexander Schätzle,
Martin Przjyaciel-Zablocki,
Thomas Hornung

dbis

Requirements

- **Study regulations**
  - Master: 16 ECTS
    → 480 hours ~ 34h/week p.P.
  - Bachelor: 6 ECTS
    → 180 hours ~ 13h/week p.P.

  - Team size: 3-4 students
  - Project report
  - Final presentation

  - Workload of every student must be clearly distinguishable

Organization

- **Time & Place:**
  - Monday 14–17 pm (c.t.)
  - Room: SR 00 007 (MMR), Building 106
- **Next Meeting:**
  - Monday, 8. November 2010 14–17 pm (c.t.)
  - Room: SR 00 007 (MMR), Building 106
- **Further Schedule:**
  - Meeting with short presentations of all groups
  - Further individual meetings upon consultation

Project Schedule

- **Induction phase**
  - Until ~~Tuesday, 2. November 2010~~ **Today!**
  - Project placing + Classification of groups
- **Short presentation**
  - 8. November 2010
  - Project introduction
  - Milestones
  - Internal work-sharing
- **Implementation phase**
  - Programming & Documentation
  - 10. / 17. January 2011: Status report of the Milestones (Meeting or Presentation)
- **Final presentation**
  - 7. February 2011
  - Contribution of project report (14. February 2011)

2. Project

- Bachelor

Motivation

- **Facebook** (2010)[1]
  - > 500 Million active users
    saving profiles, pictures, comments, news
  - > 900 Million sites, groups, events, …
  - Usage: > 700 Billion minutes per month

- **How can we handle and analyze such huge amounts of Data?**

- **Solution: Distributed-Computing?**

Source:
(1) Facebook Press Room (22.09.2010)
http://www.facebook.com/press/info.php?statistics

Task Description

- **Analysis of social Networks**
  - Query Language with navigational capabilities
  - **F**riend **o**f **a** **F**riend (FOAF) Queries
  - Search for shortest paths

- **Means of expression**
  - Startnode (e.g. „Chris“)
  - Specification of edges (Location steps) (e.g. „knows“)
  - Filter (e.g. age = 18, gender = female)
  - *Shortest Paths*

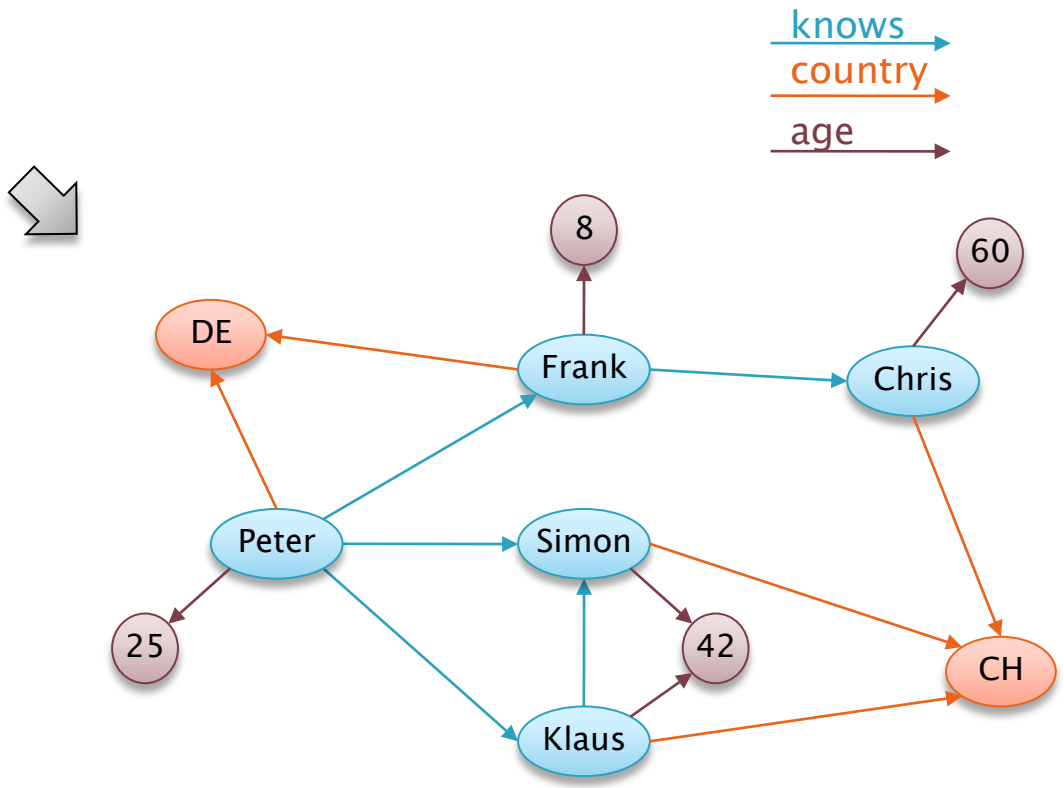
Task Description (2)

- **Data basis**
  - Graph of a social Network (Last.fm)
  - Friendship relationships
  - Properties of Users and Tracks
  - Analyzing interesting characteristics „Six-Degrees of Separation “
  - Representation: RDF-like Triples (no explicit URIs)

- **Last.fm**
  - Musikdienst mit sozialem Netzwerk
  - Freundschaften, Musiktitel, Hör-Profile, u.v.m.!

Example

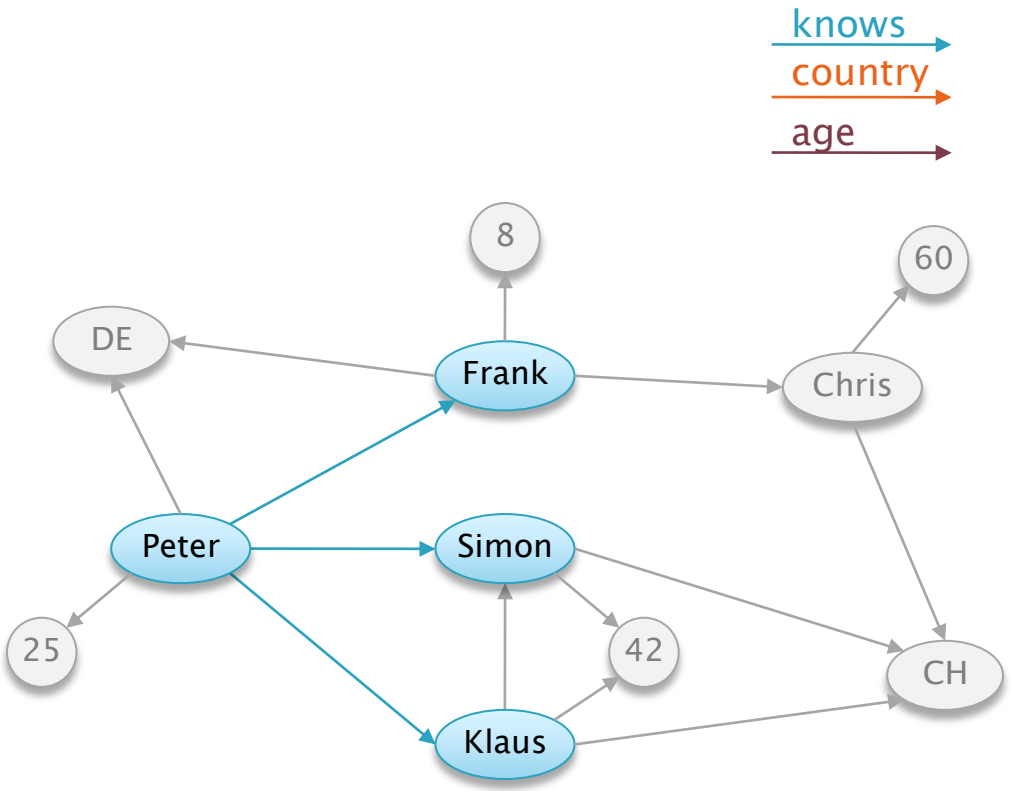
```
Peter       knows       Simon
Peter       knows       Frank
Peter       country     DE
Peter       age         25
Frank       knows       Chris
Frank       age         8
               ...
```



Example: Startnode

- `Peter :: knows.`

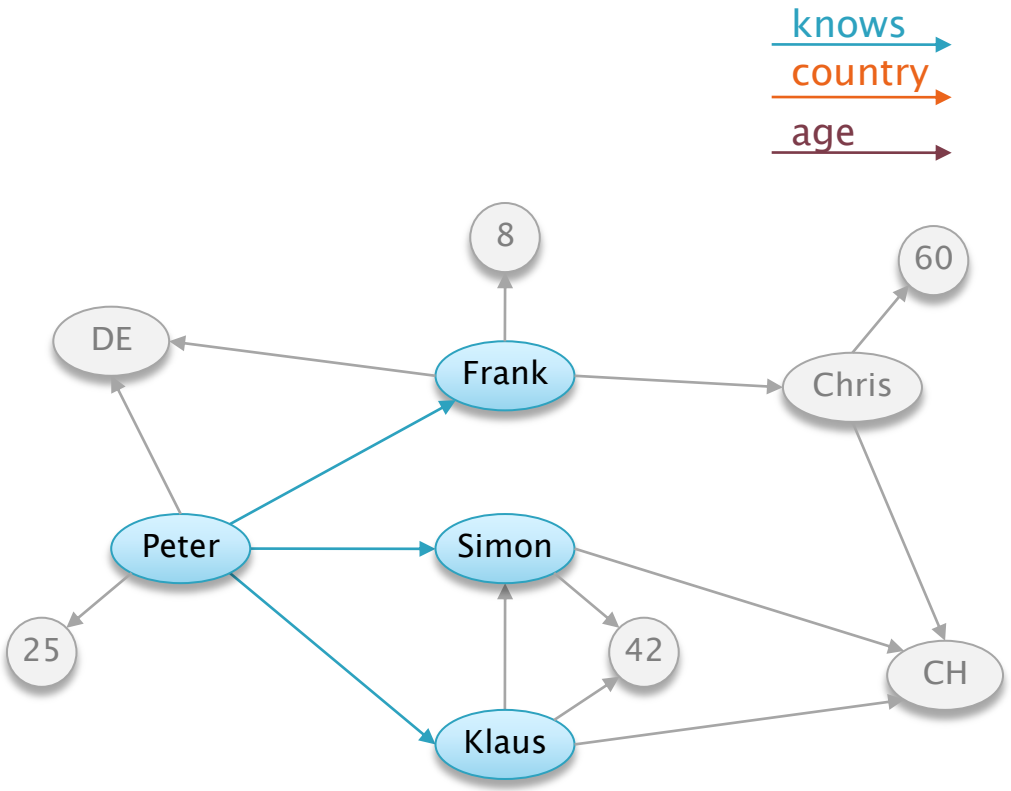
- Ergebnis
  - `Peter (knows) Frank`
  - `Peter (knows) Simon`
  - `Peter (knows) Klaus`



Example: Location Steps

- `Peter :: knows > knows > age.`

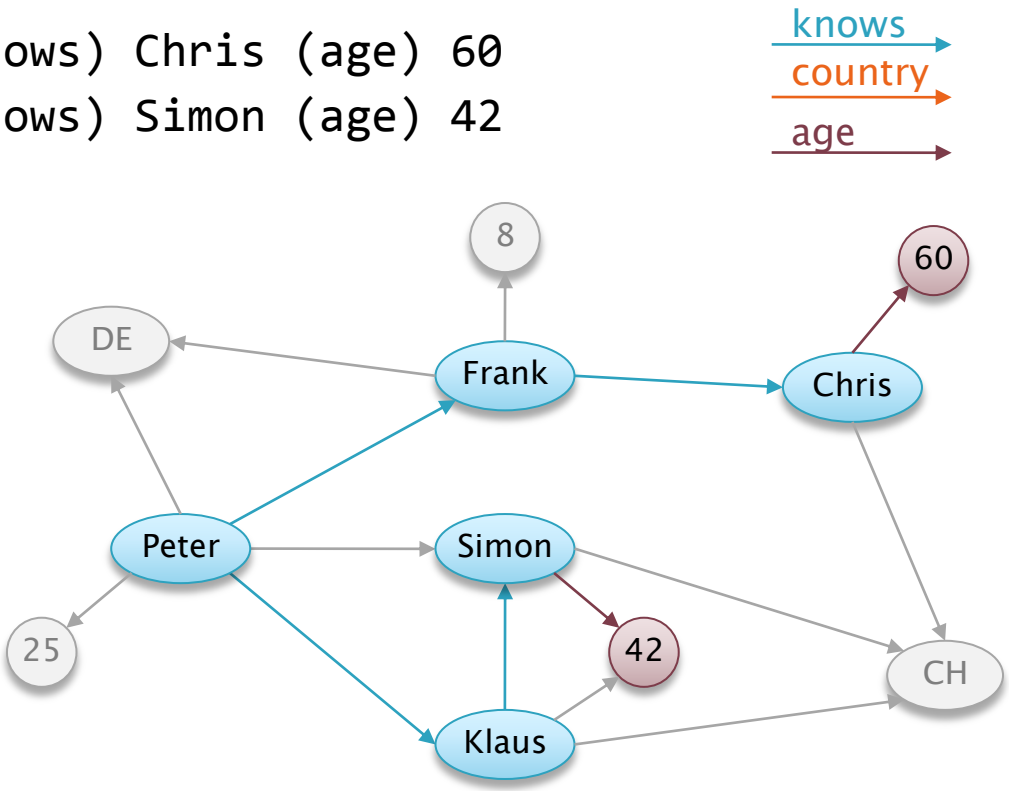
- Zwischenergebnisse
  - Peter (knows) Frank
  - Peter (knows) Klaus
  - Peter (knows) Simon



Example: Location Steps (2)

- `Peter :: knows > knows > age.`

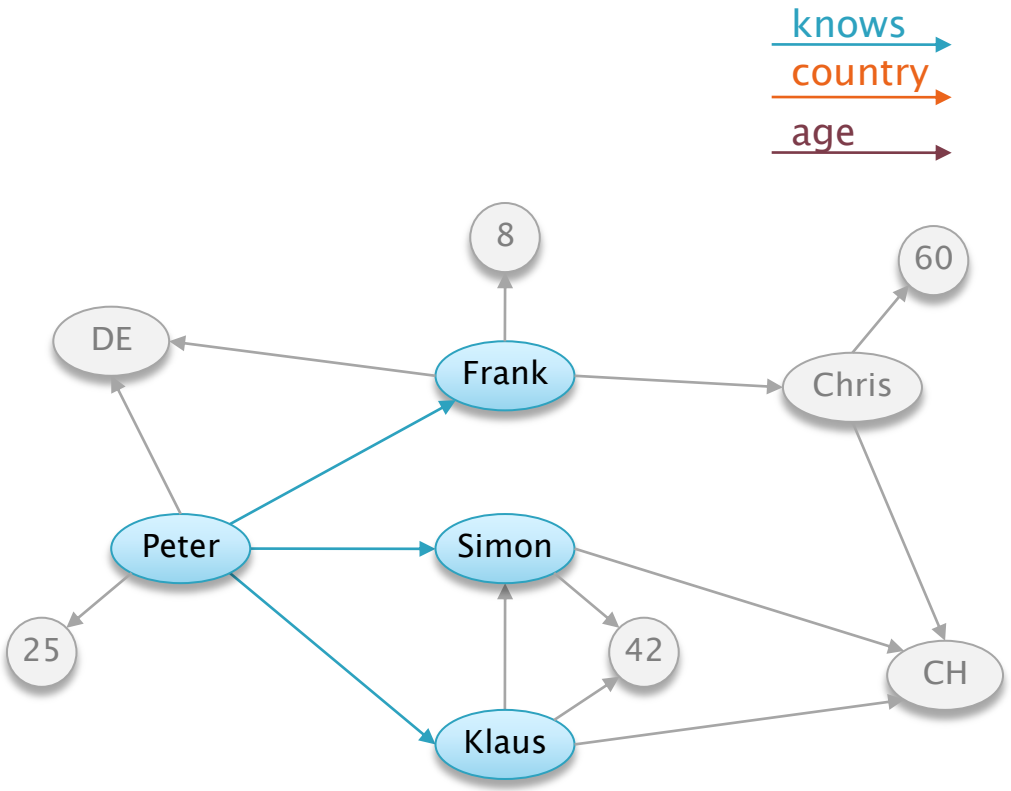
- Ergebnisse
  - `Peter (knows) Frank (knows) Chris (age) 60`
  - `Peter (knows) Klaus (knows) Simon (age) 42`
  - ~~`Peter (knows) Simon`~~



Example: Filter

- `Peter :: knows > country [equals(DE)].`

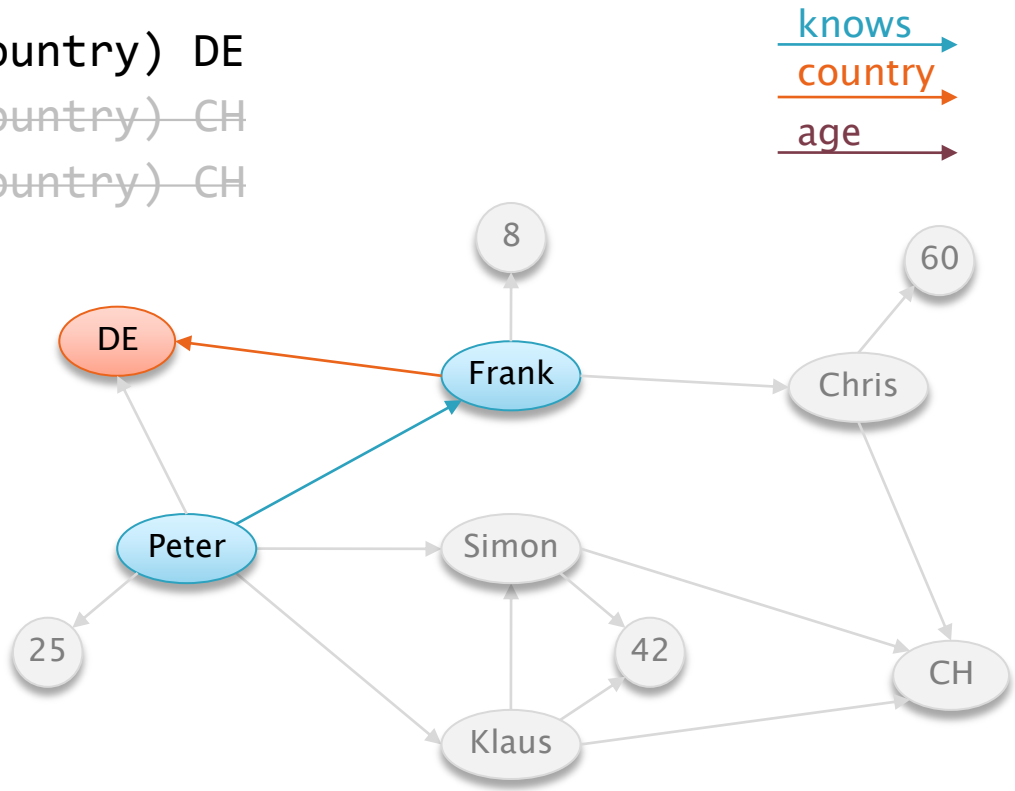
- Zwischenergebnisse
  - Peter (knows) Frank
  - Peter (knows) Klaus
  - Peter (knows) Simon



Example: Filter (2)

- `Peter :: knows > country [equals(DE)].`

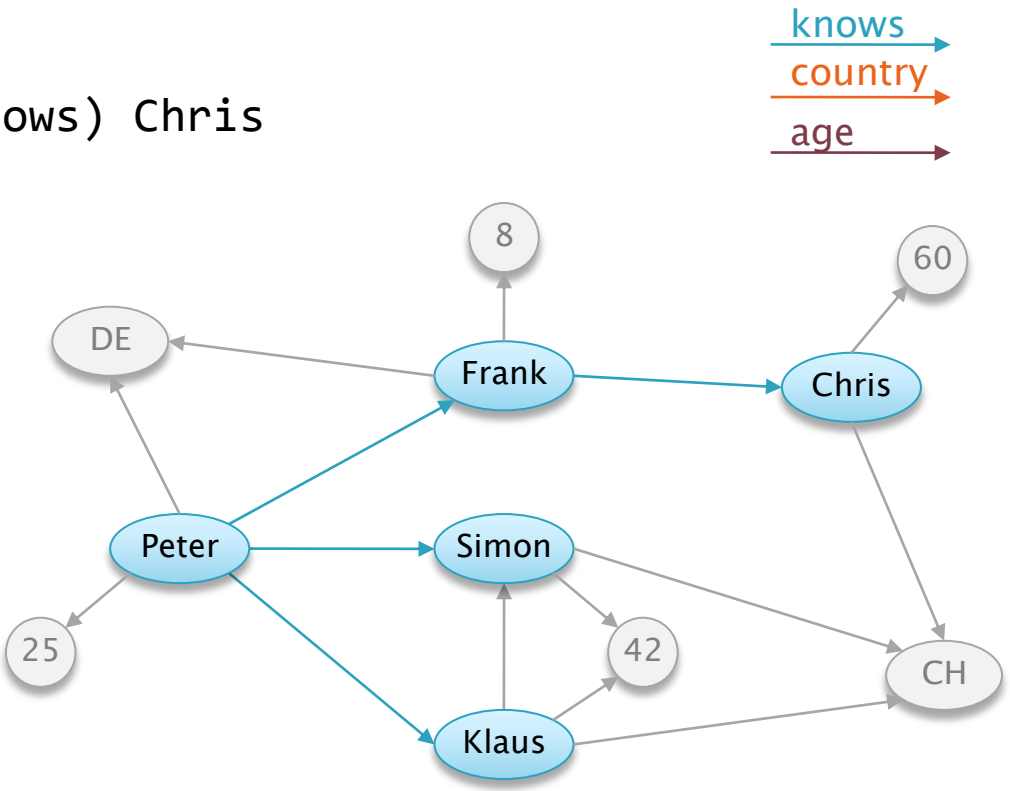
- Ergebnis
  - Peter (knows) Frank (country) DE
  - ~~Peter (knows) Klaus (country) CH~~
  - ~~Peter (knows) Simon (country) CH~~



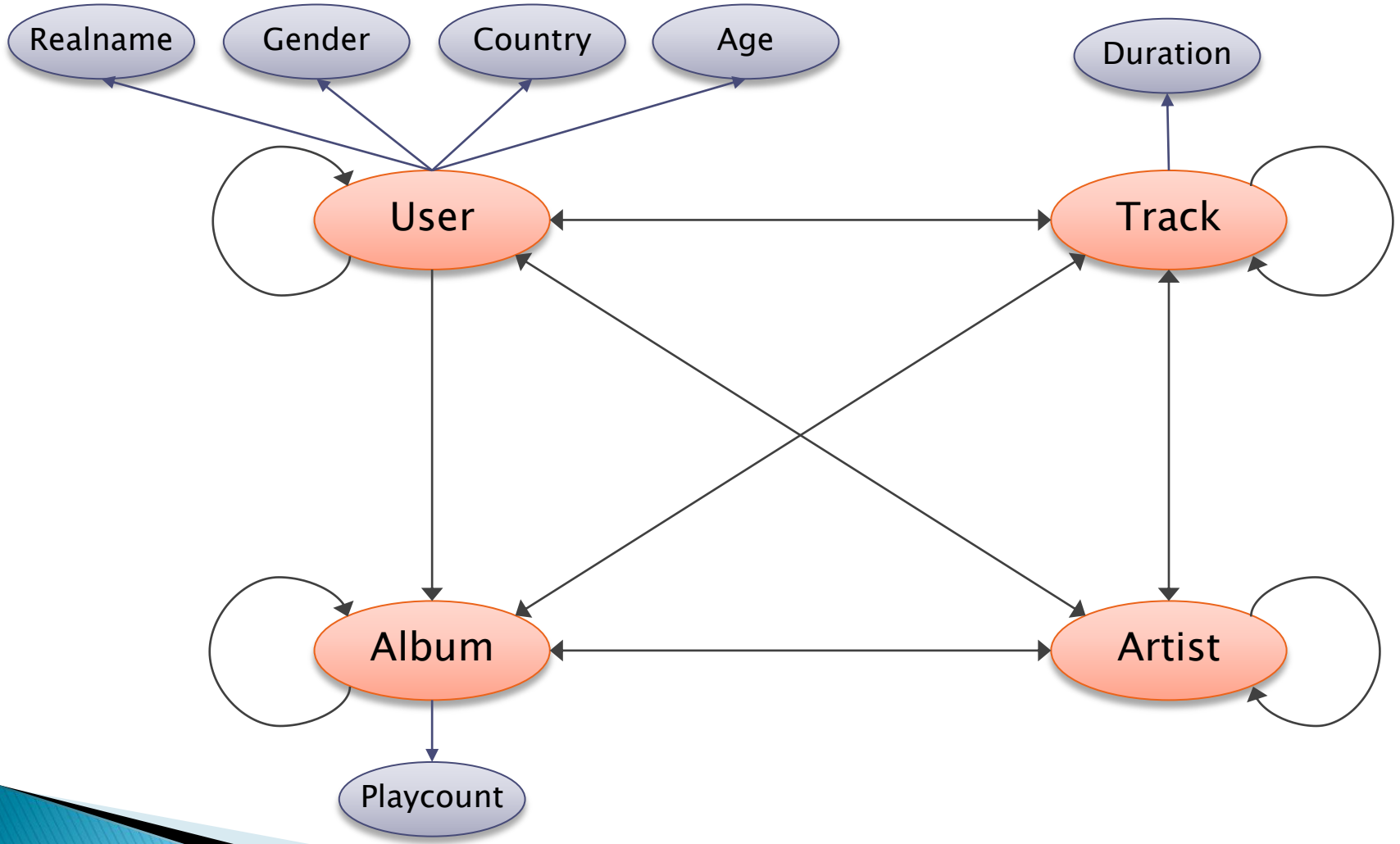
Example: *Shortest Paths*

- `Peter :: knows(*3).`

- Ergebnisse
  - Peter (knows) Frank
  - Peter (knows) Frank (knows) Chris
  - Peter (knows) Klaus
  - Peter (knows) Simon



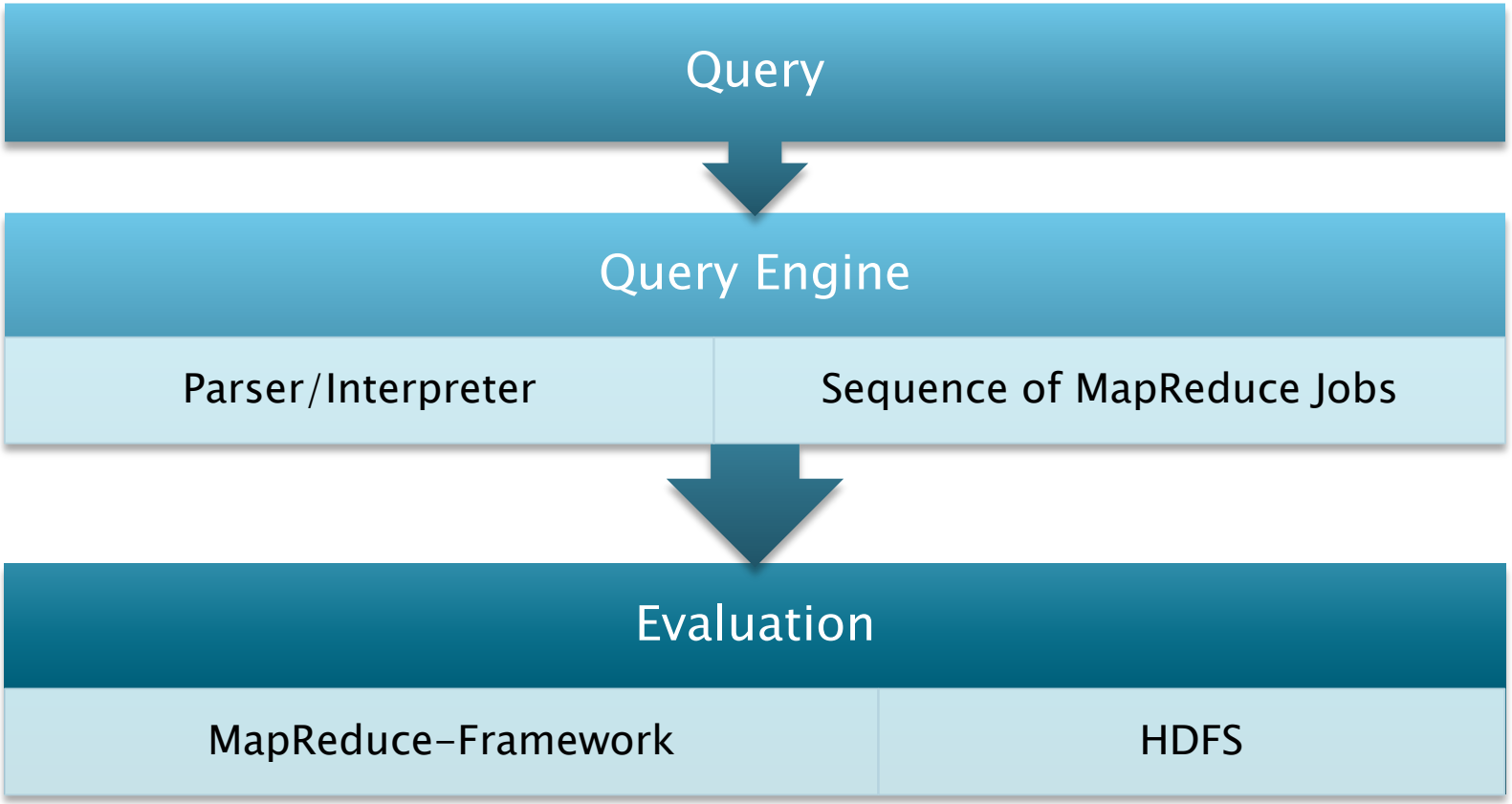
Last.fm Overview



Last.fm Overview (2)

- **User**
  - knows: 3814884
  - topArtists: 13101056
  - topTracks: 13264340
  - topAlbums: 13130619
  - listenedTo: 49830975
  - country: 180056
  - playcount: 269247
  - realname: 144040
  - gender: 268310
  - age: 153366
- **Album**
  - artist: 269900
  - tracks: 181624
  - playcount: 240347
- **Track**
  - artist: 271237
  - album: 181624
  - topFans: 8970531
  - duration: 271205
  - playcount: 271205
- **Artist**
  - tracks: 271234
  - album: 269898
  - topFans: 7439313
  - topTracks: 7728443
  - topAlbums: 936749
  - similar: 59471056

Implementation



Goal (Bachelor Projekt)

(1) Distributed Analysis of Social Networks

- Parsing of a Path Query Language
- Translating a Query in a Sequence of MapReduce-Jobs
- Storing (intermediate) results in the Cluster
- Execution of MapReduce-Jobs with Hadoop

(2) Means of Expression

- Startnode
- Multiple Location Steps
- Filter
- *Shortest Paths*

(3) Get experienced in handling MapReduce, HDFS, …

3. Teamproject

- Master

Requirements

- **Study regulations**
  - Master: 16 ECTS
    → **480 hours ~ 34h/week p.P.**
  - Recommendation: **no** parallel lectures!
  - Team size: 3-4 students
  - Project report
  - Final presentation

Requirements (2)

- Self-activated Investigation and Induction in the needed topics:
  - Triple Stores
  - SPARQL
  - Resource Description Framework (RDF)
  - MapReduce
  - Hadoop Distributed Filesystem
  - HBase
- Workload of every student must be clearly distinguishable

Task Description

- **Goal**
  - Design and Implemenation of a distributed RDF Triple Store built on top of Hadoop (MapReduce-Framework)
- **General Conditions**
  - SPARQL as Query Language (at least Basic Graph Patterns + Filter)
  - Execution in the MapReduce-Framework (Hadoop)
  - Storage strategy using HDFS or HBase

Agenda

- Now
  - Group assignment: 3-4 students per Team
  - Exchange contact informations (E-Mail, phone)
  - Plan your next team internal meeting (as soon as possible)
- Until next meeting (8. November)
  - Get to know the Project Task
  - Investigation and conceptual Design
  - Determine Milestones and Schedule (Recommendation: ~ 3 Milestones)
  - Plan internal work-sharing (regarding individual skills)
- 8. November
  - Short presentation of all groups (5-10 minutes)
  - Content: Project introduction, Milestones and internal work-sharing (perhaps overview of the planned architecture)
- Next Meeting
  - Monday, 8. November 2010 14-17 pm (c.t.)
  - Room: SR 00 007 (MMR), Building 106