

2. November 2010

(Advanced) Cloud Computing

Teamprojekt & Projekt

Veranstalter: Prof. Dr. Georg Lausen

Betreuer: Alexander Schätzle,
Martin Przyjaciel-Zablocki,
Thomas Hornung

Anforderungen

▶ Studienordnung

- **Master:** 16 ECTS
→ 480 Semesterstunden ~ 34h/Woche p.P.
- **Bachelor:** 6 ECTS
→ 180 Semesterstunden ~ 13h/Woche p.P.
- Teamgröße: 3–4 Studenten
- Projektbericht
- Abschlusspräsentation
- Arbeitsleistung einzelner Teilnehmer muss klar voneinander abzugrenzbar sein

Organisation

▶ Zeit und Ort:

- Montag 14–17 Uhr (c.t.)
- Raum: SR 00 007 (MMR), Geb. 106

▶ Nächstes Treffen:

- Montag, 8. November 2010 14–17 Uhr (c.t.)
- Raum: SR 00 007 (MMR), Geb. 106

▶ Weiterer Ablauf:

- Treffen mit Kurzpräsentationen aller Teams
- Weitere individuelle Termine auf Anfrage

Projektablauf

▶ Einarbeitungsphase

- Bis ~~Dienstag, 2. November 2010~~ **Heute!**
- Themenvergabe + Gruppeneinteilung

▶ Kurzpräsentation

- 8. November 2010
- Projektvorstellung
- Eigene Milestones
- Interne Arbeitsaufteilung

▶ Implementierungsphase

- Programmierung & Dokumentation
- 10. / 17. Januar 2011: Zwischenbericht zu den Milestones (Treffen oder Präsentation)

▶ Abschlusspräsentation

- 7. Februar 2011
- Abgabe Projektbericht (14. Februar 2011)

2. Projekt

»» Bachelor

Motivation

- ▶ **Facebook (2010)¹**
 - > 500 Millionen aktive Nutzer speichern Profile, Bilder, Kommentare, Nachrichten
 - > 900 Millionen Seiten, Gruppen, Events, ...
 - Nutzung: > 700 Milliarden Minuten pro Monat
- ▶ **Wie kann man solche riesigen Datenmengen verwalten & analysieren?**
- ▶ **Lösung: Distributed-Computing?**

Quelle:

(1) Facebook Pressebereich (22.09.2010)

<http://www.facebook.com/press/info.php?statistics>

Aufgabenstellung

- ▶ **Analyse sozialer Netzwerke**
 - Anfragesprache mit Navigationsmöglichkeit
 - Friend of a Friend (FOAF) Anfragen
 - Suche nach kürzesten Pfaden

- ▶ **Ausdrucksmöglichkeiten**
 - Startknoten (z.B. „Chris“)
 - Spezifikation von Kanten (Lokationsschritte) (z.B. „knows“)
 - Filter (z.B. age = 18, gender = female)
 - *Kürzeste Pfade*

Aufgabenstellung (2)

▶ Datengrundlage

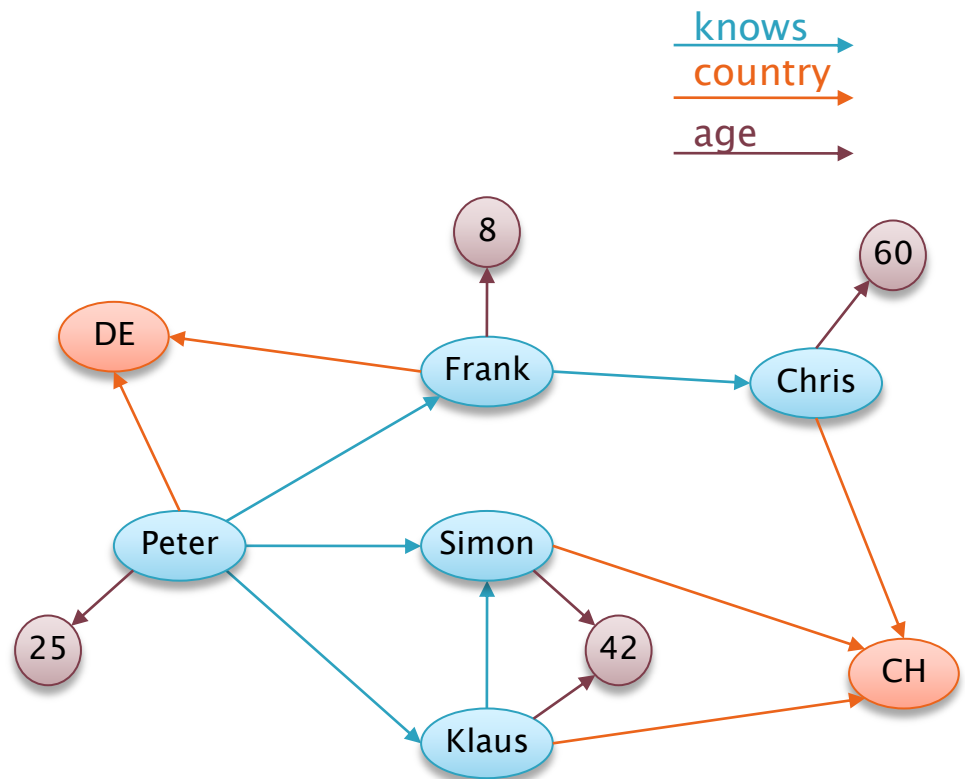
- Graph eines sozialen Netzwerkes
- Freundschaftsbeziehungen
- Eigenschaften
- Interessante Charakteristika analysieren
„Six-Degrees of Separation“
- Repräsentationsform: RDF-Graphen

▶ Last.fm

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

Beispiel

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

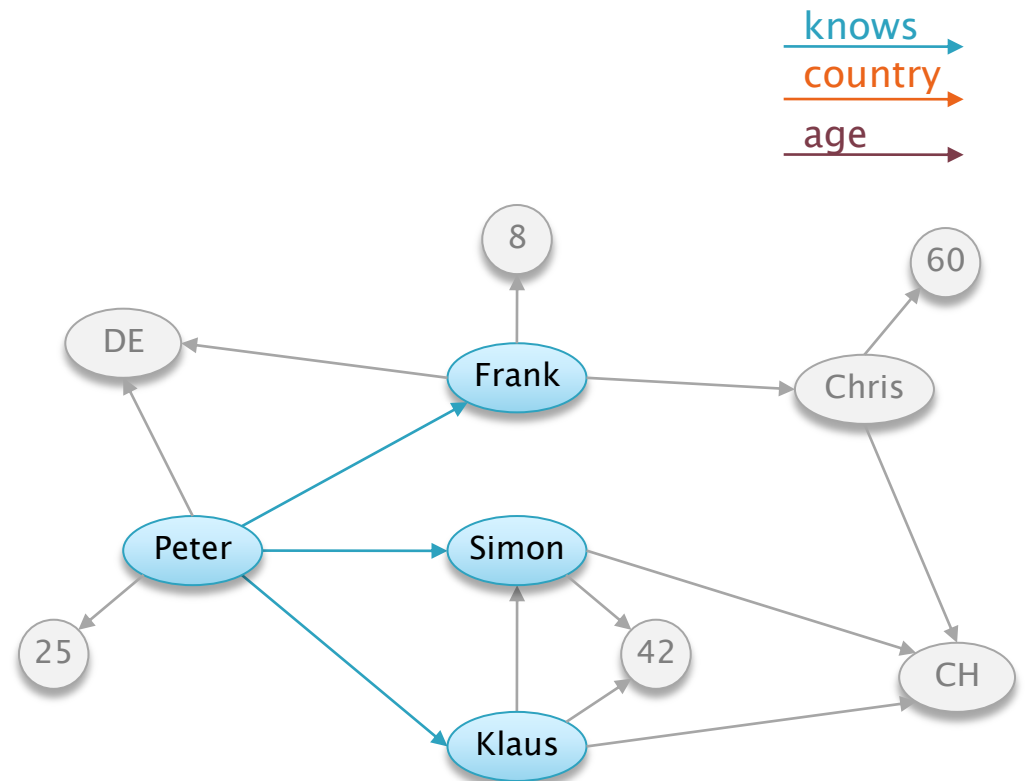


Beispiel: Startknoten

▶ Peter :: knows.

▶ Ergebnis

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

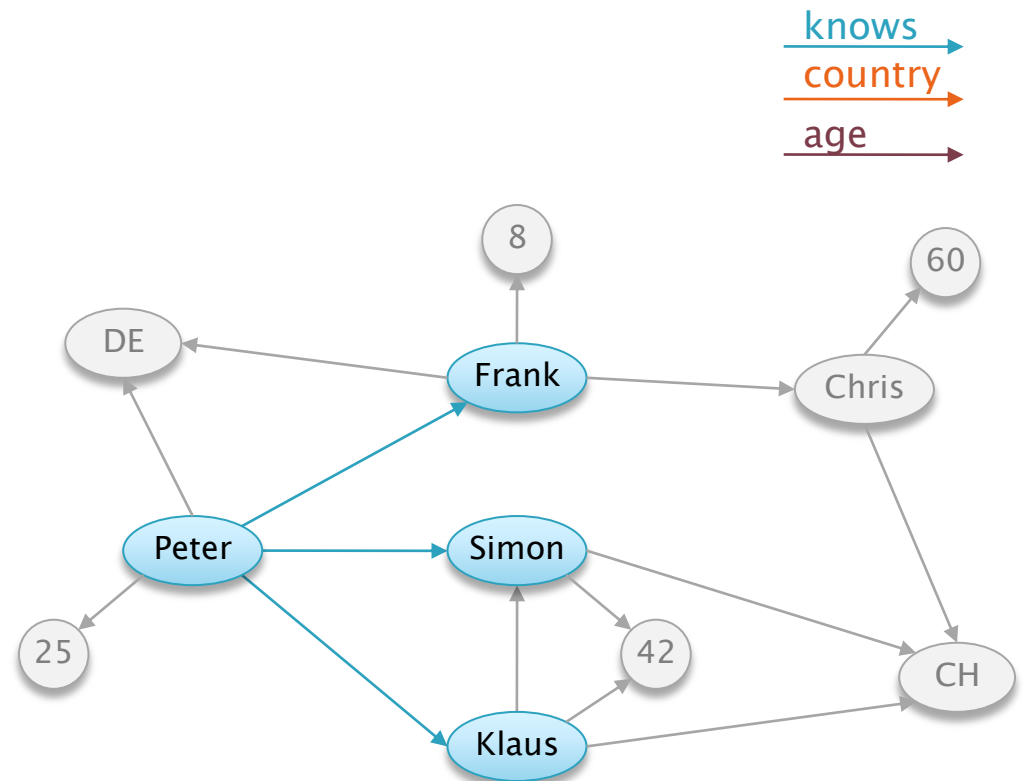


Beispiel: Lokationsschritte

▶ **Peter :: knows** > knows > age.

▶ Zwischenergebnisse

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

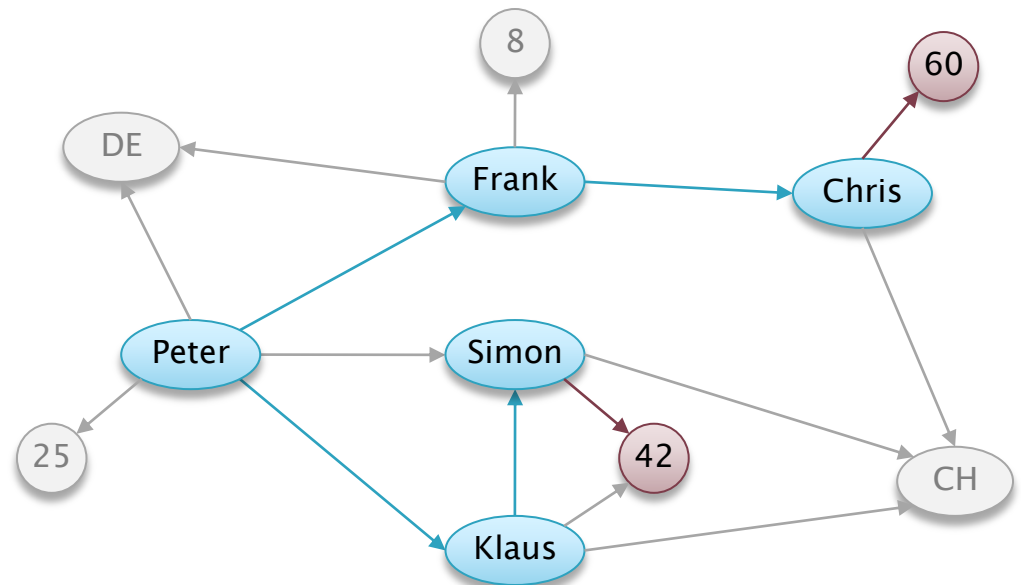
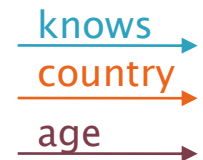


Beispiel: Lokationsschritte (2)

▶ Peter :: knows > knows > age.

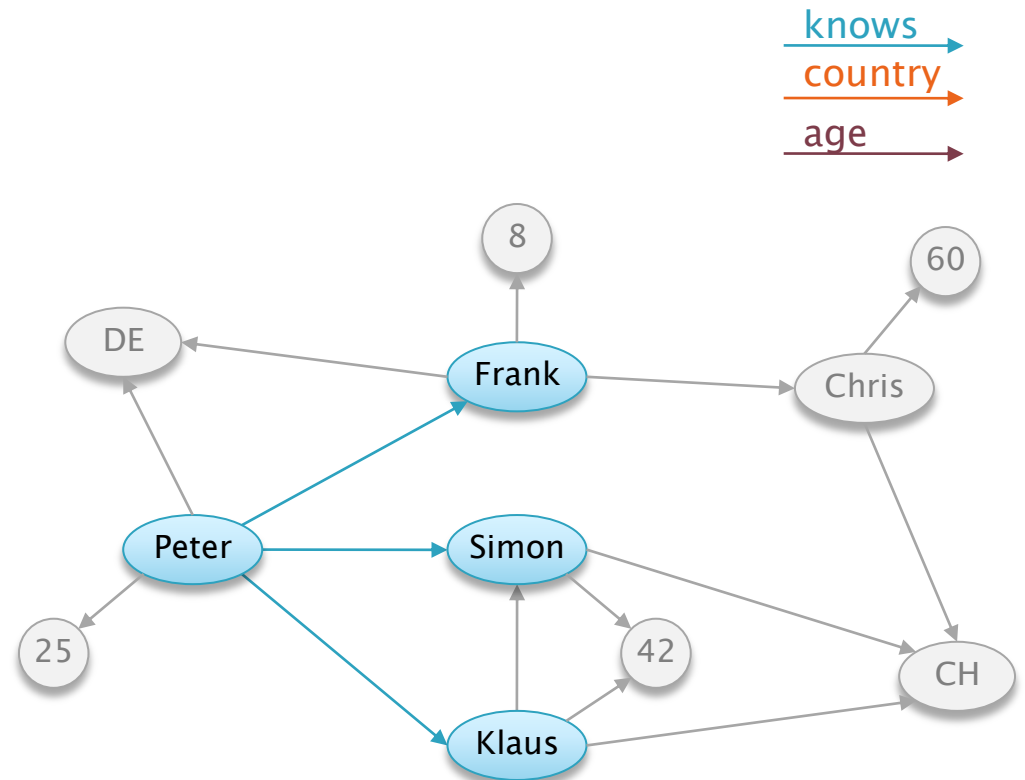
▶ Ergebnisse

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



Beispiel: Filter

- ▶ **Peter :: knows** > country [equals(DE)].
- ▶ Zwischenergebnisse
 - Peter (knows) Frank
 - Peter (knows) Klaus
 - Peter (knows) Simon

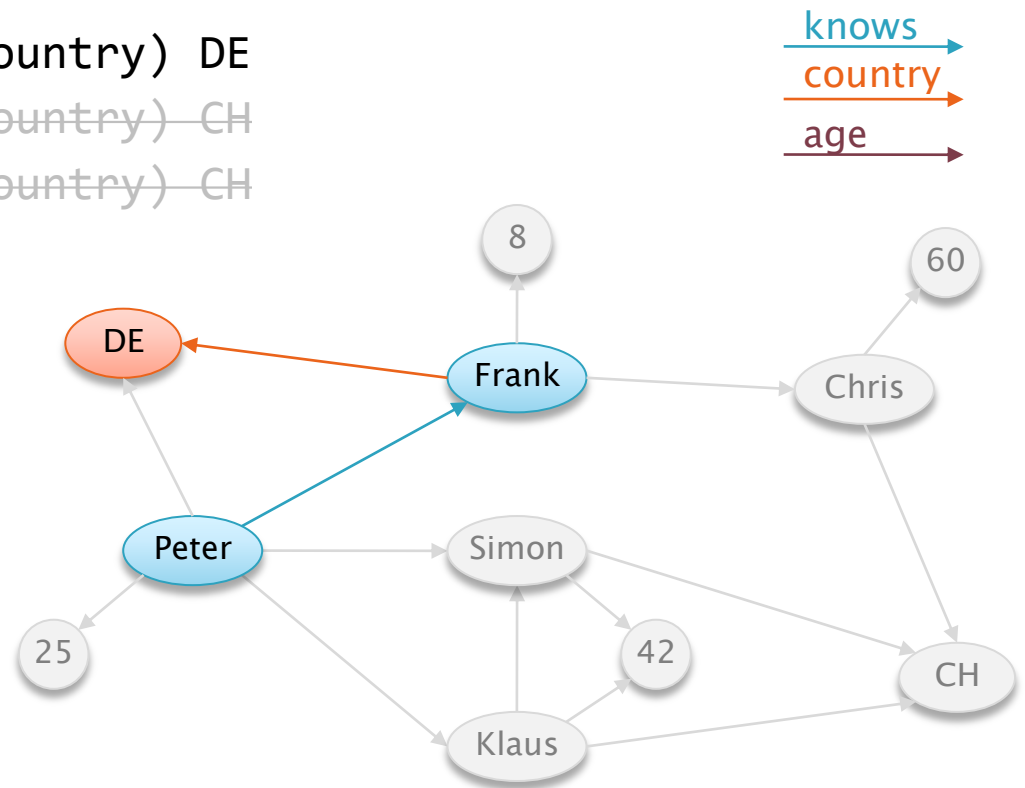


Beispiel: Filter (2)

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

▶ Ergebnis

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

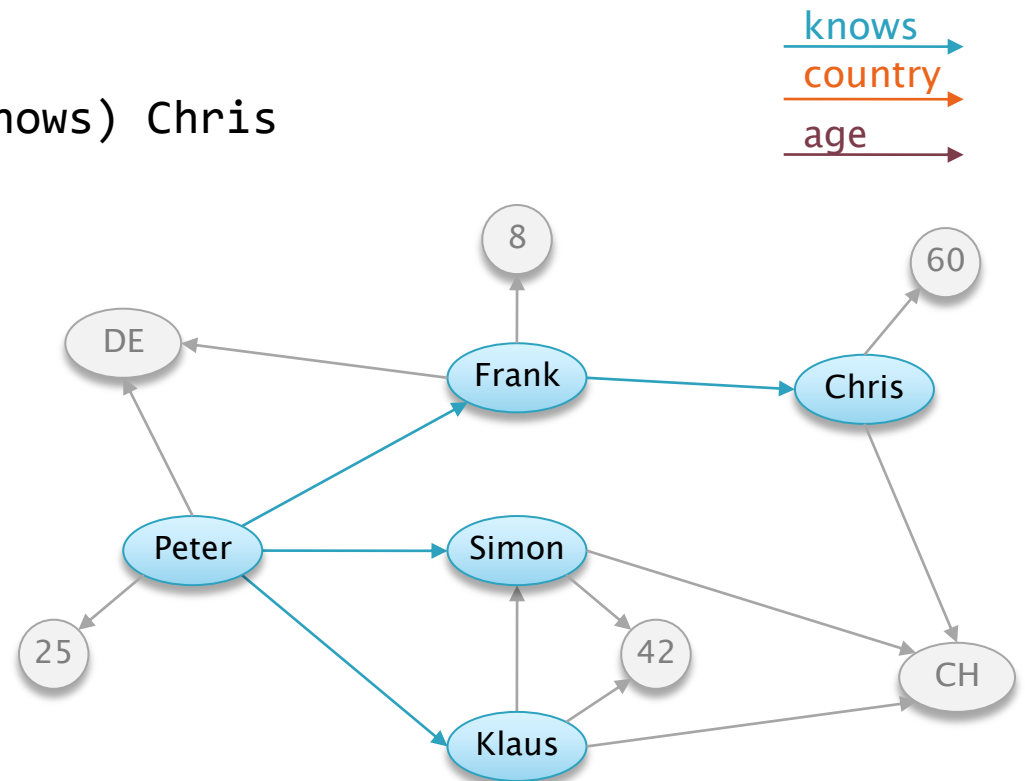


Beispiel: *Kürzeste Pfade*

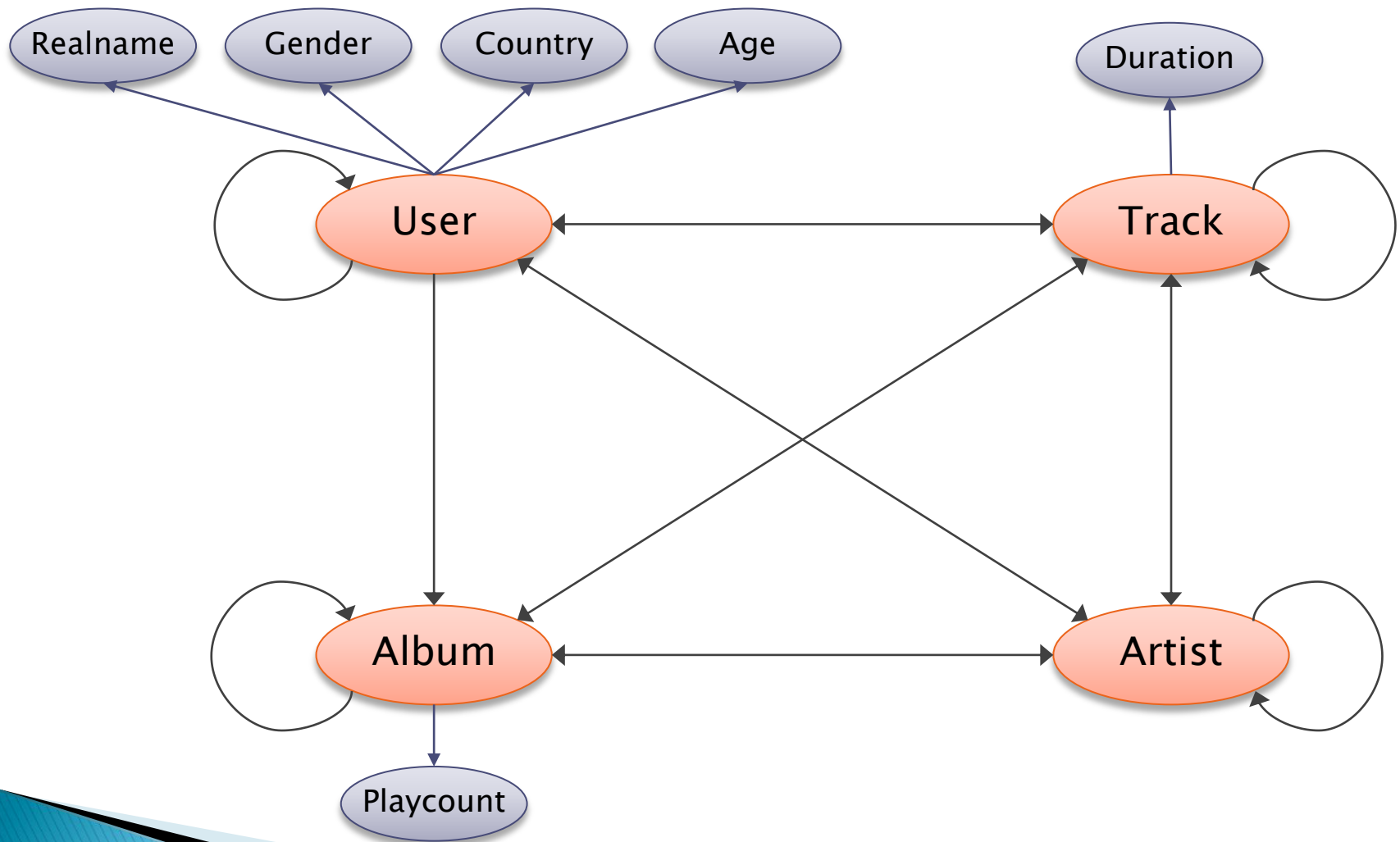
▶ Peter :: knows(*3).

▶ Ergebnisse

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



Last.fm Überblick



Last.fm Überblick (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

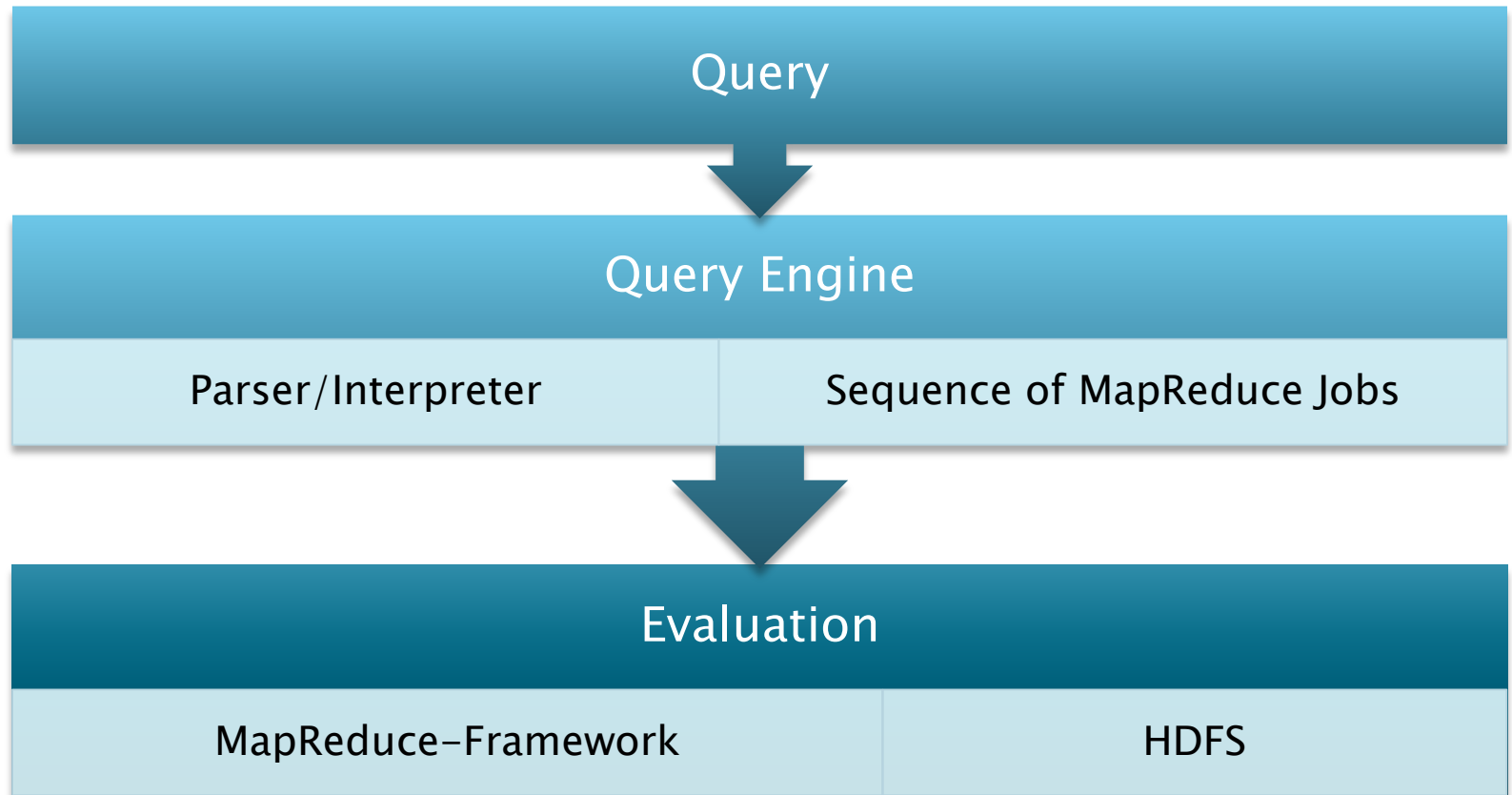
▶ Track

- artist: 271237
- album: 181624
- topFans: 8970531
- duration: 271205
- playcount: 271205

▶ Artist

- tracks: 271234
- album: 269898
- topFans: 7439313
- topTracks: 7728443
- topAlbums: 936749
- similar: 59471056

Implementierung



Zielsetzung (Bachelor Projekt)

(1) Verteilte Analyse von sozialen Netzwerken

- Parsen einer Pfadanfragesprache
- Umsetzung/Überführung der Ausdrucksmöglichkeiten in eine Sequenz an MapReduce-Jobs
- Speicherung von (Zwischen-) Ergebnissen im Cluster
- Ausführung der MapReduce-Jobs im MapReduce-Framework (Hadoop)

(2) Ausdrucksmöglichkeiten

- Startknoten
- Mehrere Lokationsschritte
- Filter
- *Kürzeste Pfade*

(3) Erfahrung im Umgang mit MapReduce, HDFS, ... sammeln

3. Teamprojekt

»» Master

Anforderungen

▶ Studienordnung

- **Master:** 16 ECTS
→ **480 Semesterstunden ~ 34h/Woche p.P.**
- Empfehlung: **keine** parallelen Veranstaltungen nebenher besuchen!
- Teamgröße: 3–4 Studenten
- Projektbericht
- Abschlusspräsentation

Anforderungen (2)

- ▶ Eigenständiges Recherchieren und Einarbeiten in die Themengebiete
 - Triple Stores
 - SPARQL
 - Resource Description Framework (RDF)
 - MapReduce
 - Hadoop Distributed Filesystem
 - HBase
- ▶ Arbeitsleistung einzelner Teilnehmer muss klar voneinander abzugrenzen sein

Aufgabenstellung

▶ Zielsetzung

- Entwurf und Implementierung eines verteilten Triple Stores (RDF–Stores) auf Basis von Hadoop (MapReduce)

▶ Rahmenbedingungen

- SPARQL als Anfragesprache (mindestens Basic Graph Patterns + Filter)
- Auswertungsprozess im MapReduce–Framework
- Speicherkonzept auf Grundlage von HDFS oder HBase

Agenda

▶ Jetzt

- Gruppeneinteilung: 3–4 Personen pro Team (in Liste eintragen)
- Kontaktinformationen (E-Mail, Telefon) austauschen
- Gemeinsame Treffen planen

▶ Aktuelle Woche

- Aufgabenstellung kennenlernen
- Recherche und konzeptueller Entwurf
- Milestones mit Zeitplan festlegen (Empfehlung: 3–4 Punkte)
- Interne Arbeitsaufteilung (bzgl. individueller Fähigkeiten)

▶ 8. November

- Kurzpräsentation aller Teilnehmer (5–10 Minuten)
- Inhalt: Projektvorstellung, Milestones und Arbeitsaufteilung

▶ Nächstes Treffen

- Montag, 8. November 2010 14–17 Uhr (c.t.)
- Raum: SR 00 007 (MMR), Geb. 106