Universität Rostock Traditio et Innovatio

Ringvorlesung: Forschung @ DBIS

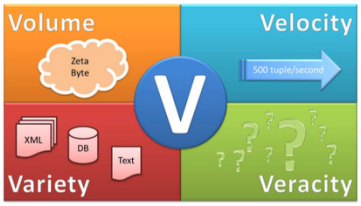
Tanja Auge

Wintersemester 16/17

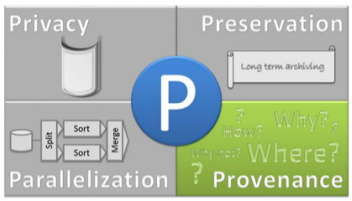
1 Die vier P

Social-Networks wie Facebook und Twitter, Forschungseinrichtungen wie das „European Council for Nuclear Research“, kurz CERN und Log-Analysen im Web oder von Sensoren produzieren eine Unmenge an zu verarbeitenden Daten. Dieses Datenvolumen, ihre große Vielfalt sowie die Geschwindigkeit der Datenverarbeitung werden hierbei unter dem Begriff Big Data zusammengefasst. Die Bezeichnung Big Data wird i.A. für Datenvolumina verwendet, die so umfangreich sind, dass sie nur schwer mit klassischen Mitteln verarbeitet werden können [11]. Zur Charakterisierung werden die folgenden vier Eigenschaften unterschieden (siehe Abbildung 1.1a):

- Data at Rest (Volume): Verwaltung großer Datensammlungen
- Data in Motion (Velocity): Unmittelbare Verarbeitung der Datenströme
- Data in Many Forms (Variety): verschiedene Datentypen
- Data in Doubt (Veracity): Unzuverlässigkeit oder Unsicherheit der Daten



(a) Die 4 V: Volume, Velocity, Variety und Veracity



(b) Die 4 P: Privacy, Preservation, Parallelization und Provenance

Abbildung 1.1: Vom Ergebnis zur Daten-Herkunft

Bei der Analyse von Big Data in Assistenzsystemen können vier Schwerpunkte gesetzt werden (siehe Abbildung 1.1b):

- Privacy: Privatheit / Datenschutz
- Preservation: Langzeitarchivierung der Daten
- Parallelization: Effiziente Analyse auf großen Datenmengen
- Provenance: Herkunft der Daten

Privacy

Das Ziel der Privacy-Analyse besteht in der Gewährleistung der Privatheit der Nutzer von Assistenzsystemen: Systeme, die im Hintergrund laufen und den Nutzer in seinen Anwendungen unterstützen, zum richtigen Zeitpunkt eingreifen und Hilfe auf akustischem oder optischem Weg anbieten. Assistenzsysteme sind vertrauenswürdig, diskret und können bei Bedarf abgeschaltet werden. Einen Abgleich der Ziele des Assistenzsystems gegen die Privatheitsansprüche des Nutzers auf Basis der Anonymisierungseigenschaften von Daten sind Gegenstand des Langzeitprojektes PArADISE (Privacy AwaRe Assistive Distributed Information System Environment) [5].

Preservation

Die Langzeitarchivierung von Daten ist ein langwieriger und zeitaufwendiger Prozess. Analoge Medien wie Bücher, Pergament oder Höhlenmalereien - um nur einige zu nennen - können eine „Haltbarkeit“ von bis zu 500 Jahren aufweisen. Eine ähnlich lange digitale Archivierung kann durch Bilder oder die Verfilmung der Daten erreicht werden. An der Universität Rostock werden diese hochvernetzten Dokumentmengen mit Hilfe von typisierten, gerichteten Hypergraphen im Rahmen des Langzeitprojektes HyDRA (a HYpergraph of Documents in a Relational Archive) modelliert. Forschungsgegenstand sind hier unter anderem die Anfrageverarbeitung und -optimierung sowie die Dokumentmodellierung und -speicherung [5].

Parallelization

Die Parallelisierung von Big Data Analytics wird ebenfalls im Projekt PArADISE untersucht. Ziel ist eine automatische Transformation in SQL, um die SQL-Anfrage anschließend zu optimieren und hochgradig zu parallelisieren. So können die Datenbankanfragen weiter optimiert und eine größere Daten- und Transaktionssicherheit erreicht werden [10].

Provenance

Das Langzeitprojekt METIS (Management, Evolution, Transformation und Integration von Schemata) vereint die Verfahren zur Integration, Transformation und Evolution von Daten. Auch die Frage der Datenherkunft wird in diesem Projekt untersucht [5]. Schwerpunkte einer Provenance-Untersuchung wären beispielsweise die Rekonstruierung einer Datenbankanfrage oder die Nachvollziehbarkeit einer (wissenschaftlichen) Aussage [7].

2 Provenance

Das Wort *provenance* stammt aus dem englischen Sprachgebrauch und bedeutet wörtlich übersetzt *Herkunft* oder *Ursprung* [3]. Im Sinne der Informatik beschäftigt sich das Provenance Management mit der Rückverfolgbarkeit eines Ergebnisses bis zu den relevanten Originaldaten. Anwendung findet dieses Management etwa in der Interpretation von Statistiken, der Analyse von Wahlergebnissen oder Umfragen sowie experimentellen Auswertungen [7].

Provenance-Anfragen und -Antworten:

Provenance-Anfragen:

- where-Anfrage: Woher kommen die Daten?
- why-Anfrage: Warum dieses Ergebnis?
- how-Anfrage: Wie kommt das Ergebnis zustande?
- why not-Anfrage: Warum fehlt ein bestimmtes Element im Ergebnis?

Provenance-Antworten [8]:

- extensionale Antwort:
  Tupel aus den Originaldaten → Antwort auf where- oder why-Anfrage
- intensionale Antwort: Beschreibung der Daten
- anfragebasierte Antwort:
  Selektionsprädikate → Antwort auf why- und how-Anfragen
- modifikationsbasierte Antwort:
  Vorschlag zur minimalen Änderung der Auswertung
  → Antwort auf why not-Anfrage

Typischerweise werden sowohl vier Provenance-Anfragen als auch vier Provenance-Antworten unterschieden. Die ersten drei Anfragen können dabei bzgl. ihres Informationsgehalt wie folgt geordnet werden: where $\subset$ why $\subset$ how. Während die where-Anfrage, im Falle einer extensionalen Antwort, lediglich die Namen der benötigten Tabellen wiedergibt, liefert die how-Anfrage eine konkrete Berechnungsvorschrift in Form eines Provenance-Polynoms [2].

Provenance-Halbringe

Ein Provenance-Halbring ist ein kommutativer Halbring $(K, +, \cdot, 0, 1)$ mit

- K: Menge der Tupelidentifikatoren (jedes Tupel der Ausgangsdatenbank hat einen eindeutigen Tupelidentifikator),
- $(K, +)$: kommutative Halbgruppe mit neutralem Element 0 (entspricht der Vereinigung oder Projektion (ohne Duplikate)),
- $(K, \cdot)$: kommutative Halbgruppe mit neutralem Element 1 (entspricht dem natürlichen Verbund),
- Distributivgesetz:

$$\forall x, y, z \in K : [(x + y) \cdot z = x \cdot z + y \cdot z] \wedge [z \cdot (x + y) = z \cdot x + z \cdot y].$$

Jedem Ergebnistupel einer Provenance-Anfrage kann ein eindeutiges Provenance-Polynom bestehend aus den Tupelidentifikatoren der Quelltupel zugeordnet werden. Diese können durch die Verwendung von K-Semimodulen um Aggregate wie MIN, MAX, AVG, SUM und COUNT erweitert werden [1].

Experimente in der Biologie [4]:

Die Anwendung von Provenance Management soll im Folgenden kurz anhand des Leibniz-Instituts für Ostseeforschung in Warnemünde erläutert werden. Durch diverse Messungen, Analysen und Simulationen werden eine Vielzahl heterogener biologischer, chemischer, physikalischer und numerischer Daten erzeugt, die anschließend im Rahmen des Forschungsdatenbankenmanagements langfristig gespeichert und verwaltet werden sollen. Fragen nach der Herkunft der Daten, ihrer Relevanz und der Rekonstruierbarkeit von Ergebnissen können nun mit Hilfe von Provenance-Fragen beantwortet werden. So kann beispielsweise bei der Messung des Sauerstoffgehalts der Ostsee (siehe Abbildung 2.1) gezielt nach dem rot eingerahmten Ausreißer gefragt werden. Die Provenance-Antwort hierauf kann sowohl intensional als auch extensional erfolgen.

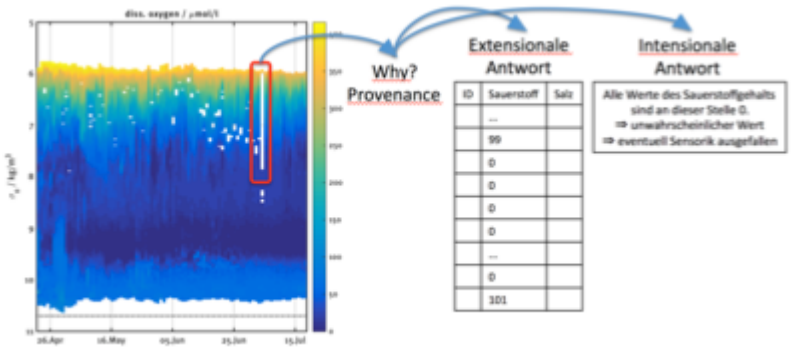


Abbildung 2.1: Why-Provenance am Beispiel des Sauerstoffgehalts der Ostsee

3 Offenes Problem

Ein unter Privacy-Aspekten gutes Assistenzsystem speichert nur Daten, die die Privatheit des Nutzers nicht verletzen. Provenance-Anfragen würden hier in der Regel eine intensionale, d.h. beschreibende, Antwort liefern; eine extensionale Antwort widerspricht größtenteils den Privacy-Aspekten. Eine automatische Rekonstruktion der Originaldaten, kann sich in solchen Fällen allerdings als schwierig herausstellen. Die hierfür nötigen inversen Schemaabbildungen werden zwar bzgl. ihrer Eigenschaften untersucht und im Bereich der Schema-Evolution erfolgreich angewendet, könnten aber auch in der Provenance-Analyse von Nutzen sein [6].

Ein offenes Problem im Bereich der Rekonstruierbarkeit besteht beispielsweise in der Berechnung dieser inversen Abbildungen durch „Rettung“ von Annotationen. Sei dazu zunächst eine Datenbank eines Fitness-Studios $d(S)$ wie folgt gegeben:

- P_1: Person (1234, Max Mustermann, Mitarbeiter)
- P_2: Person (1234, Max Mustermann, Sportler)
- K_1: Kursbelegung (1234, Bauch-Beine-Po)
- K_2: Kursbelegung (1234, Rücken-Fit)
- K_3: Kursbelegung (1234, Kettlebell)

Eine Anfrage nach den Teilnehmern des Kurses *Bauch-Beine-Po* liefert nun *Max Mustermann* sowohl als *Mitarbeiter* als auch als *Sportler* im Rahmen einer extensionalen Antwort. Formal entspricht dies den Zeugen $\{P_1, K_1\}$ und $\{P_2, K_1\}$.
Ein Zeuge ist ein Tupel in $T = \{t_1, ..., t_n\}$, der Menge der Ergebnistupeln t_i einer Anfrage q_1 auf die Quelldatenbank $d(S)$. Die why-Provenance von $t \in T$ entspricht nun der Zeugenmenge T selbst. In obigem Beispiel bedeutet dies für das Tupel $t = \{P_1, K_1\}$: $\text{why}(t) = T$, denn die Anfrage q_1 = „Teilnehmer des Kurses Bauch-Beine-Po“ liefert gerade $T = \{\{P_1, K_1\}, \{P_2, K_1\}\}$.
Rekonstruierbar ist eine Anfrage q_1 also für die von den Zeugen $T = \{t_1, ..., t_n\}$ aufgespannte Teildatenbank d_W der Quelldatenbank $d(S)$. Das METIS-Projekt verfolgt nun die Idee, die Annotationen d_A zu retten, so dass $T = q_2^{-1}(t \cup d_A)$. Für den Fall, dass die Annotationenmenge gerade der von den Zeugen aufgespannten Teildatenbanken d_W entspricht - also einer erhaltenden Transaktion - folgt die Rekonstruierbarkeit automatisch und es gilt $q_2^{-1} = q_1^{-1}$. Andernfalls handelt es sich bei q_2 um eine Pseudoinverse und es stellt sich die Frage, nach einer stärkeren Einschränkung von d_A?

Betrachten wir erneut das obige Beispiel erweitert um die folgenden Personen und Kursbelegungen:

- P_3 : Person (1235, Paul Paulchen, Mitarbeiter)
- P_4 : Person (1236, Peter Petersen, Sportler)
- P_5: Person (1237, Simon Simonson, Mitarbeiter)
- P_4: Kursbelegung (1235, Rücke-Fit)
- K_5: Kursbelegung (1236, Bauch-Beine-Po)
- K_6: Kursbelegung (1237, Bauch-Beine-Po).

Die Anfrage $q_3 =$ „Wie viele Mitarbeiter nehmen am Bauch-Beine-Po-Kurs teil?“ liefert die Zahl 2. Dabei kan die Annotationenmenge wie folgt reduziert werden:

- P_1: Person (1234, Max Mustermann, Mitarbeiter) $\Rightarrow P_1' =$ (a, Mitarbeiter)
- P_5: Person (1237, Simon Simonson, Mitarbeiter) $\Rightarrow P_5' =$ (b, Mitarbeiter)
- K_1: Kursbelegung (1234, Bauch-Beine-Po) $\Rightarrow K_1' =$ (a)
- K_6: Kursbelegung (1237, Bauch-Beine-Po) $\Rightarrow K_6' =$ (b)

Die exakte Identifikationsnummer ist eventuell nicht relevant oder darf aufgrund von Privacy-Aspekten nicht veröffentlicht werden; sie kann aber durch eine eindeutige Variable ersetzt werden. Der Name des Mitarbeiters ist für die obige Anfrage irrelevant und der Kursname bereits in der Anfrage enthalten. Die Annotationenmenge kann also auf $d_A = \{P_1', P_5', K_1', K_6'\}$ reduziert werden.

Literaturverzeichnis

[1] Y. Amsterdamer, D. Deutch, V. Tannen: *Provenance for Aggregate Queries*, Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 153–164, **2011**

[2] T. Auge, S. Brossmann, P. Wilsdorf: *Endpräsentation ProSA-Gruppe*, Neueste Entwicklungen in der Informatik, Präsentation, **SoSe 16/17**

[3] P. Hemetsberger : Deutsch-Englisch-Wörterbuch, `http://www.dict.cc/?s=provenance`, **07.03.2017**

[4] A. Heuer : *Digitale Bibliotheken und Multimedia-Retrieval*, Lehrstuhl für Datenbank- und Informationssysteme, Vorlesungsskript, **WiSe 16/17**

[5] A. Heuer : *Langzeitprojekte*, Lehrstuhl für Datenbank- und Informationssysteme, `https://dbis.informatik.uni-rostock.de/forschung/langzeitrahmenprojekte/`, **2017**

[6] A. Heuer : ***METIS** in PArADISE: Provenance Management bei der Auswertung von Sensordatenmengen für die Entwicklung von Assistenzsystemen*, Datenbanksysteme für Business, Technologie und Web (BTW 2015) - Workshopband, 2.-3. März 2015, Hamburg, Germany, **2015**

[7] A. Heuer : *Neueste Entwicklung in der Informatik - Motivation*, Lehrstuhl für Datenbank- und Informationssysteme, Vorlesungsskript, **SoSe 16**

[8] A. Heuer : *Neueste Entwicklung in der Informatik - Provenance*, Lehrstuhl für Datenbank- und Informationssysteme, Vorlesungsskript, **SoSe 16**

[9] A. Heuer : *Theorie relationaler Datenbanken*, Lehrstuhl für Datenbank- und Informationssysteme, Vorlesungsskript, **SoSe 16**

[10] A. Heuer : *Ringvorlesung: Forschung @ DBIS*, Lehrstuhl für Datenbank- und Informationssysteme, Vorlesungsskript, **WiSe 16/17**

[11] Innovation is freedom : OVH Big Data, `https://www.ovh.de/dedicated_server/HG/big-data-glossar.xml`, **05.03.2017**