Technischer Bericht des Instituts für Informatik der Universität Rostock: CS-04-15, ISSN: 0944-5900
erhältlich unter http://www.ls-dbis.de/digbib/dbis-tr-cs-04-15

Das PArADISE-Projekt

Big-Data-Analysen für die Entwicklung von Assistenzsystemen

Andreas Heuer
Lehrstuhl Datenbank- und Informationssysteme
Institut für Informatik
Universität Rostock
18051 Rostock, Deutschland
heuer@informatik.uni-rostock.de

Holger Meyer
Lehrstuhl Datenbank- und Informationssysteme
Institut für Informatik
Universität Rostock
18051 Rostock, Deutschland
hme@informatik.uni-rostock.de

ZUSAMMENFASSUNG

Bei der Erforschung und systematischen Entwicklung von Assistenzsystemen fallen eine große Menge von Sensordaten an, aus denen Situationen, Handlungen und Intentionen der vom Assistenzsystem unterstützten Personen abgeschätzt (modelliert) werden müssen. Neben Privatheitsaspekten, die bereits während der Phase der Modellbildung berücksichtigt werden müssen, sind die *Performance* des Analysesystems sowie die *Provenance* (Rückverfolgbarkeit von Modellierungsentscheidungen) und die *Preservation* (die langfristige Aufbewahrung der Forschungsdaten) Ziele unserer Projekte in diesem Bereich. Speziell sollen im Projekt PArADISE die Privatheitsaspekte und die Performance des Systems berücksichtigt werden. In einem studentischen Projekt wurde innerhalb einer neuen *experimentellen* Lehrveranstaltung im reformierten Bachelor- und Master-Studiengang Informatik an der Universität Rostock eine Systemplattform für eigene Entwicklungen geschaffen, die auf Basis von klassischen zeilenorientierten Datenbanksystemen, aber auch spaltenorientierten und hauptspeicheroptimierten Systemen die Analyse der Sensordaten vornimmt und für eine effiziente, parallelisierte Verarbeitung vorbereitet. Ziel dieses Beitrages ist es, die Ergebnisse dieser studentischen Projektgruppe vorzustellen, insbesondere die Erfahrungen mit den gewählten Plattformen PostgreSQL, DB2 BLU, MonetDB sowie R (als Analysesystem) zu präsentieren als auch die Erfahrungen mit dieser Art von experimenteller Lehrveranstaltung im Kontext der Bologna-Regelungen weiterzugeben.

ACM-Klassifikation

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation*; K.4.1 [**Computers and Society**]: Public Policy Issues—*privacy*

Stichworte

Assistenzsysteme, Big Data Analytics, Spaltenorientierte und hauptspeicheroptimierte Datenbanksysteme

Ein zweiseitiges Extended Abstract dieses Beitrags erscheint im Tagungsband des 27th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 26.05.2015 - 29.05.2015, Magdeburg, Germany.

1. EINLEITUNG

Ein Forschungsschwerpunkt am Institut für Informatik der Universität Rostock ist die Erforschung und systematische Entwicklung von Assistenzsystemen. Da in Assistenzsystemen unterstützte Personen durch eine Vielzahl von Sensoren beobachtet werden, müssen bei der datengetriebenen Modellierung von Situationen, Handlungen und Intentionen der Personen aus großen Datenmengen mittels Machine-Learning-Methoden entsprechende Modelle abgeleitet werden: ein Performance-Problem bei einer Big-Data-Analytics-Fragestellung.

Da Personen *beobachtet* werden, müssen auch Privatheitsaspekte bereits während der Phase der Modellbildung berücksichtigt werden, um diese bei der konkreten Konstruktion des Assistenzsystems automatisch in den Systementwurf zu integrieren. Somit gibt es für die Datenbankforscher unter anderem die Teilprobleme der performanten Berechnung der Modelle als auch der Wahrung der Privatheitsansprüche des Nutzers, die zu lösen sind und die in einer langfristigen Projektgruppe des Datenbanklehrstuhls angegangen werden: im Projekt **PArADISE** (Privacy AwaRe Assistive Distributed Information System Environment) werden effiiziente Techniken zur Auswertung von großen Mengen von Sensordaten entwickelt, die definierte Privatheitsansprüche der späteren Nutzer per Systemkonstruktion erfüllen.

Während wir in [Heu15] ausführlicher auf die Verknüpfung der Aspekte *Privatheit* (Projekt PArADISE) und *Provenance* (Projekt METIS) eingegangen sind, werden wir uns in diesem Beitrag auf die beiden Schwerpunkte des PArADISE-Projektes konzentrieren, das ist neben der Privatheit die *Performance* durch Parallelität und Verteilung.

Im Folgenden werden wir im Abschnitt 2 kurz die Architektur von Assistenzsystemen einführen, wobei wir den Schwerpunkt auf die Phase der Situations-, Aktivitäts- und Intentionserkennung legen. Danach werden wir die Erforschung und Entwicklung von Assistenzsystemen als ein wissenschaftliches Experiment ansehen, in dem Forscher eine große Anzahl von Sensordaten auswerten müssen und aus ihnen Situations-, Aktivitäts- und Intentionsmodelle entwickeln (Abschnitt 3.2). Im Abschnitt 4 werden wir die Grundlagenforschungsthemen für unsere Arbeitsgruppe definieren, die von Performance und Privatheit über das Provenance Management bis zur Langzeitarchivierung (Preservation) reichen. Speziell die Aspekte der Performance und Privatheit bearbeiten wir im Projekt PArADISE, das im Abschnitt 5

Abbildung 1: Pyramidenarchitektur von Assistenzsystemen

eingeführt wird. Die Ergebnisse der ersten Implementierungen und ein Vergleich verschiedener zeilen- oder spaltenorientierter DBMS als Basis für die Auswertung werden in Abschnitt 6 vorgestellt. Der Artikel endet mit den Erfahrungen aus einer experimentellen Lehrveranstaltung, in der die erste Phase dieses Projektes mit einer studentisches Projektgruppe umgesetzt wurde (Abschnitt 7).

2. ASSISTENZSYSTEME

Ähnlich einem menschlichen Assistenten soll ein Assistenzsystem mich unterstützen, im Hintergrund arbeiten (ambient), mich nicht stören, zum richtigen Zeitpunkt eingreifen und Hilfe anbieten (diese in üblichen Fällen auf optischem oder akustischem Wege), vertrauenswürdig und diskret sein und sich bei Bedarf abschalten lassen.

Um seine Assistenzaufgaben zu erfüllen, besteht ein Assistenzsystem üblicherweise aus fünf Schichten (siehe Abbildung 1 nach [HKHT06], auch in [HKG14]). Dabei deutet die Pyramidenform an, dass in der untersten Schicht ständig viele Daten (etwa von Sensoren) erzeugt werden, in der obersten Schicht aber nur im Bedarfsfall (also eher selten) ein akustischer oder optischer Hinweis, also eine geringe Datenmenge, ausgegeben wird.

Sensoren in der Umgebung der Person sollen Situation und Tätigkeit der Person erfassen, um ihr assistieren zu können. **Ortungskomponenten** sollen die genaue Position der Person bestimmen, etwa zur Detektion dementer Patienten mit Weglauftendenzen. Sensoren und Ortungskomponenten befinden sich in der Umgebung, in benutzten Geräten oder am Körper der Person (Armband, Brille, . . .).

Damit ein Assistenzsystem seine Aufgabe erfüllen kann, müssen verschiedene (heterogene) Geräte in der Umgebung der Person **vernetzt** und zur Erreichung des Assistenzziels **spontan gekoppelt** werden.

Sensordaten müssen gefiltert, erfasst, ausgewertet, verdichtet und teilweise langfristig verwaltet werden. Aufgrund der extrem großen Datenmenge (Big Data) muss die **Verarbeitung verteilt** erfolgen: teilweise eine Filterung und Verdichtung schon im Sensor, im nächsterreichbaren Prozessor (etwa im Fernseher oder im Smart Meter in der Wohnung) und im Notfall über das Internet in der Cloud. Neben Daten des Assistenzsystems müssen auch fremde Daten etwa über das Internet berücksichtigt werden, beispielsweise Wartungspläne beim Auto oder die elektronische Patientenakte beim Patienten. Allgemein können hier natürlich auch die Daten sozialer Netzwerke, Kalenderdaten der Nutzer oder Wettervorhersage-Daten ausgewertet werden, falls sie für das Assistenzziel eine Rolle spielen.

Nach der Auswertung der Sensordaten erfolgt die **Situations-, Handlungs- und Intentionserkennung** (siehe den folgenden Abschnitt 3.2 für die Entwicklung diesbezüglicher Modelle).

Schließlich erfolgt die **multimodale Interaktion** mit dem Nutzer. Üblich sind optische Signale wie eine Meldung auf dem Fernseher oder über eine Warnlampe, sowie akustische Signale wie eine Ansage über ein Radio oder Alarmtöne über einen Lautsprecher. Wenn der assistierten Person ein Signal nichts mehr nützt, kann auch eine Meldung an Angehörige, Ärzte oder Notfallzentralen ausgelöst werden. Umgekehrt kann man dem Assistenzsystem auch selbst Hinweise geben, etwa über Touchscreen, Gesten oder Bewegungen oder mittels Sprache (Kommandos).

3. BIG DATA ANALYTICS

Eine Kernaufgabe bei der Erforschung und Entwicklung ist die datengetriebene Modellierung von Situationen, Handlungen und Intentionen, die eine Fragestellung im Forschungsgebiet Big Data Analytics sind. Wir führen hier zunächst unsere *technische* Definition von *Big Data* ein, bevor wir die konkreten Analysefragestellungen erläutern.

3.1 Big Data

Big Data [Mar15] ist ein derzeitiges Hype-Thema nicht nur in der Informatik, das in seiner technischen Ausprägung auf vielfältige Forschungsprobleme führt. Im Gegensatz zu [Dit15] werden wir das Thema Big Data somit nicht politisch (NSA, Snowden, Überwachung) verstehen, sondern technisch, wobei wir auf Privatheitsaspekte als Schwerpunkt unserer Forschung noch eingehen werden.

Von der technischen Seite sind Big-Data-Probleme mit den vier *V* charakterisiert:

- *Data at Rest (Volume)*: Es sind Terabytes bis Exabytes an Daten zu verwalten und zu analysieren.
- *Data in Motion (Velocity)*: Die Daten sind Stromdaten, bei der Verarbeitung muss trotzdem in kürzester Zeit reagiert werden. Schnelle Filtermechanismen auf eintreffenden Daten sind notwendig.
- *Data in Many Forms (Variety)*: Es gibt heterogene Daten jeglicher Art: strukturierte, semistrukturierte, unstrukturierte Daten wie Text, Bild, Video, Audio.
- *Data in Doubt (Veracity = Correctness)*: Die zu verarbeitenden Daten sind ungenau, weil teilweise fehlend, mehrdeutig, die Daten treffen teilweise zu spät ein (Latenz), müssen approximiert werden. Inkonsistenzen etwa durch Heterogenitäten müssen bereinigt werden.

Wichtig wegen *Volume* und *Velocity* ist das Prinzip des *Process and Forget*: Die Filtermechanismen können nur einen (definierten) Bruchteil der Daten langfristig speichern, was eine Herausforderung für das Provenance Management sein wird.

Big Data Analytics ist nun das Problem komplexer Analysen auf diesen Daten. In Datenbankbegriffen sind diese komplexen Analysen iterative Anfrageprozesse.

3.2 Big Data Analytics bei der Entwicklung von Assistenzsystemen

Aufgrund der Sensor- und Ortungsdaten sowie der weiteren über das Internet erhältlichen Daten muss das Assistenzsystem eine Situations- und Handlungserkennung vornehmen sowie eine Handlungsvorhersage (Intentionserkennung), um proaktiv eingreifen zu können.

Die Situation ist dabei die aktuelle Umgebungsinformation, die Handlung das, was die Person, der assistiert wird, gerade durchführt. Die Intentionserkennung oder Handlungsvorhersage muss voraussagen, was die Person in Kürze tun wird.

Die Handlungs- und Intentionserkennung ist ein aktueller Forschungsgegenstand der Informatiker an der Universität Rostock [KNY+14], etwa im DFG-Graduiertenkolleg MuSAMA. Dabei erheben die Forscher in langen Versuchsreihen eine extrem hohe Anzahl von Sensordaten, aus denen sie mit diversen Analyseverfahren die entsprechenden Modelle ableiten. Diese Analyseverfahren sind — wenn sie ohne Datenbankunterstützung auf Dateisystemen mit Analysewerkzeugen wie R ausgeführt werden — mehrwöchige Prozesse.

Der Umfang der Daten soll hier kurz an einem Szenario abgeschätzt werden: Die Erfassung eines kurzen Handlungsablauf von 40 Minuten einer Versuchsperson durch einen Motion-Capturing-Anzug, EMG-Instrumentierung (Elektromyografie) mit paralleler Video-Aufzeichnung erzeugt einen Rohdatenbestand von 10 GByte. Werden für die Entwicklung eines Handlungsmodells, das den Alltag repräsentiert, Daten von vierzehn Versuchspersonen über jeweils etwa 100 Stunden aufgezeichnet, ergibt dieser Versuch Daten im Umfang von 14 Terabyte, die für Analysen verfügbar sein müssen, um entsprechende Modelle zu rechnen.

Ziel der Forscher ist neben der Modellbildungen für Handlung und Intention die Erkenntnis, wie die große Anzahl von Sensoren im Versuch für den praktischen, späteren Einsatz des Assistenzsystems drastisch eingeschränkt werden kann, ohne die Vorhersagequalität zu mindern. Für die Ableitung dieser Informationen müssen unter anderem alle Analysefunktionen invertiert werden, um die für die Modellbildung entscheidenden Anteile der Originaldaten zu finden. Letzteres ist auch ein Problem im Provenance Management, das die Experimentverläufe mit den Ergebnisableitungen begleiten soll. Eine Einschränkung sowohl der Anzahl der Sensoren als auch der Menge und Granularität der erfassten Daten ist auch aus einem anderen Grund wichtig: sie kann die Privatheitsanforderungen der Nutzer des Assistenzsystems realisieren helfen.

4. DIE VIER P ZU DEN VIER V

Die Forschungsschwerpunkte der Rostocker Datenbankgruppe lassen sich in diesem Zusammenhang mit vier *P* charakterisieren, die im Folgenden näher erläutert werden sollen.

Forschung und Entwicklung: In der Forschungs- und Entwicklungsphase eines Assistenzsystems ist das vorrangige Ziel, eine effiziente Modellbildung auf großen Datenmengen zu unterstützen. Dabei sollte möglichst automatisch eine Selektion der Daten (Filterung wichtiger Sensordaten nach einfachen Merkmalen) und eine Projektion der Daten (die Beschränkung der großen Sensormenge auf wenige, besonders aussagekräftige Sensoren) vorgenommen werden. Die nötige Effizienz in dieser Phase führt auf unser Forschungsthema **P3: Performance**. Da während der Entwicklung bei fehlerhafter Erkennung von Handlungen und Intentionen die dafür zuständigen Versuchsdaten ermittelt werden müssen, führt die Rückverfolgbarkeit der Analyseprozesse in der Entwicklung auf unsere Forschungsthemen **P2: Provenance Management** und **P4: Preservation** (Langfristarchivierung von Forschungsdaten).

Einsatz: In der Einsatzphase eines Assistenzsystems sind dagegen Privatheitsansprüche vorherrschend, die im Gesamtsystem durch stufenweise Datensparsamkeit erreicht werden können (unser Forschungsthema **P1: Privatheit**, Privacy). Eine weitere Verdichtung (auch Reduktion und Aggregation) der live ausgewerteten Daten unterstützen aber nicht nur die Privatheit, sondern auch die Performance.

Die vier *P* behandeln wir in drei langfristigen Forschungsprojekten:

- *P1: Privacy* ist ein Kernthema des Projektes **PArADISE**. Die Sicherung der Privatheit durch Datensparsamkeit wollen wir unter anderem durch Ausnutzung des Big-Data-Analytics-Prinzips *Process and Forget* erreichen: in einer *vertikalen Architektur* wollen wir die großen Datenmengen möglichst sensornah verarbeiten und die Analysealgorithmen möglichst von der Cloud auf Prozessoren in der Wohnung der Person beziehungsweise direkt am Sensor vorverarbeiten und vorverdichten. Dabei rechnen wir damit, dass auf jeder Stufe von der Cloud bis zum Sensor die Leistungsfähigkeit der Rechnerknoten, die Zwischenspeichergröße und der Operatorumfang abnimmt.

- *P2: Provenance* ist ein Thema des Projektes **METIS**, in dem wir allgemein Schema-Instanz-Abbildungen und ihre Eigenschaften betrachten. Unter anderem wollen wir aufgrund von vorgenommenen Analysen und Analyseergebnissen durch inverse Schema-Instanz-Abbildungen grundlegende Charakteristika von Originaldaten wiedergewinnen [Heu15]. Neben der Feststellung der Herkunft der Daten soll auch die Plausibilität, Nachvollziehbarkeit und Rekonstruierbarkeit von Analyseergebnissen in verschiedenen Qualitätsstufen ermöglicht werden.

- *P3: Performance (Parallelität)* ist das zweite Kernthema des Projektes **PArADISE**. Ziel ist die effiziente Analyse von großen Datenmengen durch Ausnutzung hochparalleler Systeme. Dabei wollen wir Analyseprozesse, die iterativ in R formuliert wurden, automatisch auf SQL abbilden und diese iterativen Anfrageprozesse dann auf Rechnercluster (ähnlich zu Map-Reduce) verteilen. Dazu werden wir die oben erwähnte vertikale Architektur noch durch eine *horizontale Architektur* der Verteilung ergänzen.

- *P4: Preservation*: Die Langzeitarchivierung der zugrundelegenden, für die Entwicklung wichtigen Primärdaten ist ein Kernthema des Projektes **HyDRA**. Ziel ist die unverfälschte Nutzbarmachung und Rekonstruierbarkeit von Analyseergebnissen auch in einigen Jahrzehnten, über die gesetzlichen Vorschriften hinaus. Zusätzlich betrachten wir auch die Nachhaltigkeit der eingesetzten Analyse- und Zugangs-Software (im sogenannten Rostocker Modell mit dem IT- und Medi-

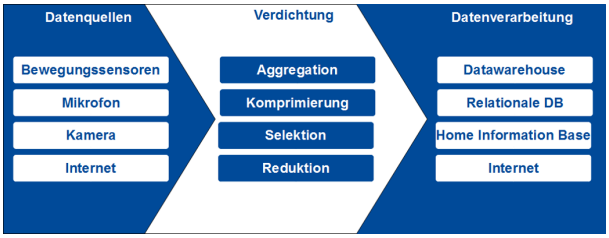

Abbildung 2: Verdichtungsmethoden für Sensordaten

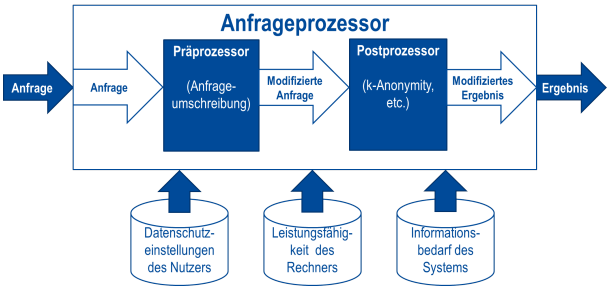

Abbildung 3: Anfragetransformationen zur Berücksichtigung von Privatheitsansprüchen des Nutzers und Informationsbedarf des Systems

enzentrum und der Universitätsbibliothek als Partner [HMB14]).

In diesem Beitrag befassen wir uns nun speziell mit dem PArADISE-Projekt.

5. DAS PARADISE-PROJEKT

Die Entwickler der Analysewerkzeuge müssen ihr Assistenzziel und die notwendigen Sensordaten zur Erreichung des Ziels und zur grundlegenden Situations-, Handlungs- und Intentionserkennung formulieren. Diese Zielformulierung wird dann in Anfragen auf Datenbanken transformiert. Weiterhin können die **Privatheitsansprüche** des Nutzers vordefiniert oder von jedem Nutzer selbst individuell verschärft werden. Auch diese Privatheitsansprüche werden in Anfragen (Sichten) auf Datenbanken umgesetzt. Durch Abgleich des Informationsbedarfs des Assistenzsystems und der Privatheitsansprüche des Nutzers kann dann die Datenbankkomponente des Assistenzsystems entscheiden, wie die Menge an Sensordaten selektiert, reduziert, komprimiert oder aggregiert werden muss, um beiden Parteien im System gerecht zu werden [Gru14]. Abbildung 2 zeigt die möglichen Datenquellen, die Verdichtungsmethoden aufgrund vom Matching von Privatheitsanforderungen und Assistenzzielen sowie die mögliche Datenverarbeitung in verschiedenen Zielsystemen (nach Grunert [Gru14]).

Ein entscheidendes Kriterium für die Vertrauenswürdigkeit eines Assistenzsystems ist noch die Frage, wie nah am Sensor die Daten bereits reduziert und verdichtet werden können: Wenn der Sensor so intelligent ist, dass er bestimmte Filtermechanismen von Datenbanksystemen beherrscht, so kann dieser bereits eine Vorfilterung vornehmen. Nur die für das Assistenzziel unabdingbaren Daten, die die Privatheit des Nutzers nicht verletzen, können dann im Rahmen des Cloud Data Management des Anbieters der Assistenzfunktionalität entfernt und verteilt gespeichert werden.

Im Projekt **PArADISE** (Privacy AwaRe Assistive Distributed Information System Environment) arbeiten wir derzeit an Techniken zur Auswertung von großen Mengen von Sensordaten, die definierte Privatheitsansprüche der späteren Nutzer per Systemkonstruktion erfüllen.

Ein erster Prototyp ist von einer studentischen Arbeitsgruppe erstellt worden. Derzeit können Analysen zur Modellbildung auf Sensordaten in SQL-92, SQL:2003 oder iterativen Ansätzen über SQL-Anweisungen realisiert und auf die Basissysteme DB2 (zeilenorientiert oder spaltenorientiert: DB2 BLU), PostgreSQL (zeilenorientiert) sowie MonetDB (spaltenorientiert und hauptspeicheroptimiert) abgebildet werden. Details dazu finden sich im folgenden Abschnitt 6. In nächster Zeit wird eine automatische Abbildung der R-Analysen unserer Assistenzsystem-Forscher auf diese SQL-Lösungen umgesetzt.

Die Privatheitsansprüche werden dabei durch automatische Anfragetransformationen abgebildet (nach [Gru14], siehe Abbildung 3). Sowohl die Privatheitsansprüche des Nutzers (Datenschutzeinstellungen wie k-Anonymität und schützenswerte Informationen oder Handlungen) als auch der Informationsbedarf des Assistenzsystems sollen dabei aufeinander abgeglichen werden (was auf ein *Query-Containment-Problem* [Chi09] führt). Später soll noch die Leistungsfähigkeit der berechnenden Prozessoren (am Sensor oder am Server) berücksichtigt werden, um über Vorfilterungen und Verdichtungen etwa direkt am Sensor entscheiden zu können.

Wie oben bereits in Abschnitt 4 erwähnt, unterscheiden wir zwischen der Erforschungs- und Entwicklungsphase des Assistenzsystems (mit Testnutzergruppen und eingeschränkter Privatheit, aber voller Provenance) und der Einsatzphase des Assistenzsystems (mit realen Nutzern, voller Umsetzung der Privatheitsansprüche, aber eingeschränkter Provenance). Allerdings sollen bereits in der Entwicklungsphase des Systems die späteren Privatheitsansprüche konstruktiv durch passende Anfragetransformationen in das Ziel-Assistenzsystem implantiert werden. Während es bei der Entwicklung von Handlungsmodellen für Assistenzsysteme um *Big Data* im eigentlichen Sinn geht, ist im Ziel-Assistenzsystem für eine Person (in einer Wohnung, im Auto) zwar immer noch eine große Menge von Stromdaten aus Sensoren vorhanden, die aber in einschlägigen Artikeln auch als *small data, where* $n = me$ [Est14] bezeichnet wird: hier ist also die Anonymität der Person nicht mehr zu verstecken, aber die Herausgabe von schützenswerten Informationen über diese Person.

6. VERGLEICH VON ZEILEN- UND SPALTENORIENTIERTEN DBMS

Während die grundlegenden Forschungsarbeiten zu PArADISE durch zwei Stipendiaten des Graduiertenkollegs MuSAMA (Hannes Grunert und Dennis Marten) in 2013 und 2014 starteten, wurden die ersten softwaretechnischen Umsetzungen des Projektes durch eine studentische Projektgruppe im Wintersemester 2014/2015 vorgenommen (siehe etwa [WKL+15]).

Eine erste Zielsetzung war dabei, grundlegende Analyse-

verfahren für Big Data und ihre Effizienz zu betrachten. Als Basis wurden drei DBMS ausgesucht: PostgreSQL, MonetDB und IBM DB2. Die komplexen Fragestellungen der Assistenzsystemforscher wurden zunächst auf zwei statistische Kernprobleme, die Korrelation (für Vorhersagemodelle) aus dem Projekt PageBeat [FBH+14] und die Regression (für Handlungsmodelle) aus dem Graduiertenkolleg MuSAMA, zurückgeführt. Diese beiden Verfahren sollten dann auf den drei Plattformen umgesetzt werden.

In der studentischen Projektgruppe wurden dann verschiedene SQL-Anfragen und R-Programme zur Lösung der Regressions- und Korrelationsprobleme entwickelt, wobei als Vorgabe (zum Vergleich) folgende fünf Stufen realisiert werden sollten:

1. Umsetzung von Regression und Korrelation in Standard-SQL-92 (also per Hand, da keine Analysefunktionen außer den klassischen Aggregatfunktionen wie COUNT, SUM und AVG vorhanden).
2. Umsetzung in SQL:2003 mit den entsprechenden OLAP-Funktionen.
3. Umsetzung mit rekursivem oder iterativem SQL, sofern in den Systemen möglich.
4. Eine Integration der SQL-Anfrage mit R-Auswertungen.
5. Eine R-Auswertung pur ohne Kopplung an SQL.

Für die Tests wurden drei virtuelle Server aufgesetzt, wobei alle drei Server mit identischer virtueller Hardware versehen wurden: 2.8 GHz Octacore CPU, 4 GB Arbeitsspeicher mit CentOS release 6.6. Die Versionen der installierten Datenbanksysteme waren:

1. DB2: v10.1.0.0 (noch als zeilenorientierte Architektur)
2. PostgreSQL: 9.4.1 (zeilenorientierte Architektur)
3. MonetDB: v1.7 (Oct2014-SP2) (eine Column-Store-Architektur, hauptspeicheroptimiert)

Da DB2 für die Projektgruppe nur in einer zeilenorientierten und nicht hauptspeicheroptimierten Version zur Verfügung stand, wurde in einer parallel laufenden Bachelor-Arbeit von Eric Rath dieselbe Fragestellung auf IBM DB2 BLU (Column Store) getestet, wobei sich gegenüber den u.a. Ergebnisse für DB2 in diesem Fall keine grundlegenden Verbesserungen ergeben haben.

Als ein Beispiel der Testergebnisse wollen wir hier für die Regressionsanalyse auf MuSAMA-Daten die Performance-Resultate angeben (siehe Abbildung 4). Die Testdaten umfassten dabei 4000 Tupel zu 666 Spalten, die von einer Bewegungsanalyse aus einem Motion-Capturing-Anzug stammten. Als iterative Komponente sollte dabei ein *Sliding Window* über 5er-Gruppen von Daten realisiert werden, wozu eine Rekursion in SQL oder alternativ eine per Hand programmierte Iteration verwendet wurde.

In Abbildung 4 ist zu sehen, dass die in MuSAMA bisher verwendete Lösung mit *Plain R* die schlechteste Effizienz aufwies, auch wenn man den Prozess des initialen Ladens der Daten in den Hauptspeicher herausrechnet. Unter den Varianten mit einer Analyse in reinem SQL-92 (Regression per Hand mit Aggregatfunktionen umgesetzt) war die MonetDB-Lösung etwas besser als die DB2-Variante, PostgreSQL fiel stärker ab. Die SQL:2003-Lösung konnte in MonetDB mangels vorhandener OLAP- und Rekursions-Fähigkeiten nicht umgesetzt werden, DB2 war hier wiederum deutlich besser als PostgreSQL. Weiterhin bemerkt man im Vergleich von SQL-92 und SQL:2003, dass der Optimierer von DB2 als auch PostgreSQL die direkte Verwendung der OLAP-Funktionen belohnt. Die beste Performance aller Varianten erreichte jedoch MonetDB mit integrierten R-Funktionen.

Im Sommersemester wird eine neue studentische Projektgruppe in einer experimentellen Lehrveranstaltung das Projekt fortsetzen. Insbesondere werden dann die vertikalen und horizontalen Verteilungen und die Anfragetransformationen für die Privatheitsaspekte betrachtet. Weiterhin werden wir eine stärkere Kopplung von SQL und R vornehmen, um hier die Effizienz zu steigern. Als vierte Plattform wollen wir zum Vergleich auch Apache Flink (auch unter Stratosphere bekannt [Mar15]) einsetzen.

7. DAS PROJEKT ALS EXPERIMENTELLE LEHRVERANSTALTUNG

Das studentische Projekt wurde als neue Form einer Projektveranstaltung durchgeführt, die im Sommer und im Winter unter jeweils wechselnden Vorzeichen angeboten wird. Dabei sollen generische Module, die in verschiedenen Phasen des Studiums im Bachelor oder im Master gewählt werden können, zu einer Veranstaltungsform zusammengeführt werden. In unserem Fall betraf dies zwei verschiedene Projektmodule im Bachelor und eine experimentelle Lehrveranstaltung mit Seminar- oder Projektcharakter im Master.

Durch den Bologna-Prozess sind solche integrierenden, virtuellen Veranstaltungen, die mehrere generische Module aus verschiedenen Studiengängen zusammenführen, kaum möglich. Insbesondere gibt es Regeln, dass keine Veranstaltung gleichzeitig als Bachelor- und als Master-Veranstaltung in einem konsekutiven Studiengang angeboten werden darf. Weiterhin kann die Hochschul-Verwaltungs-Software nicht mehrere, unterschiedliche Veranstaltungen zur selben Zeit im selben Raum anbieten: hier wird eine Integritätsprüfung zu viel durchgeführt.

Wir bieten nun in jedem Semester ein Leitmodul für einen der Studiengänge an, das wir dann als Modul eines anderen Studienganges als Ausnahmeregelung anerkennen können. Dadurch ist es möglich, Studenten vom 3. Semester des Bachelor-Studiums bis zur Master-Arbeit in einer Projektgruppe zu sammeln und zusammenarbeiten zu lassen, aber die geleisteten Arbeiten als ganz verschiedene Modulleistungen je nach Bedarf des Studenten anrechnen zu können. Alle betroffenen Module haben sechs Leistungspunkte, so dass zumindest von dieser Seite her die Kompatibilität gewahrt bleibt.

Diese virtuelle Integration von Projekt-Veranstaltungen ist der Versuch, ein sehr altes Konzept von studentischen Projektgruppen trotz Bologna-Überregulierung als sinnvolles Studienelement wieder aufleben zu lassen.

8. DANKSAGUNGEN

Wir danken der studentischen Projektgruppe PArADISE im Wintersemester 2014/2015, die im Rahmen einer experimentellen Projekt-Lehrveranstaltung die Basis für die softwaretechnische Umsetzung des PArADISE-Projektes gelegt hat: Pia Wilsdorf, Felix Köppl, Stefan Lüdtke, Steffen Sachse, Jan Svacina, Dennis Weu.

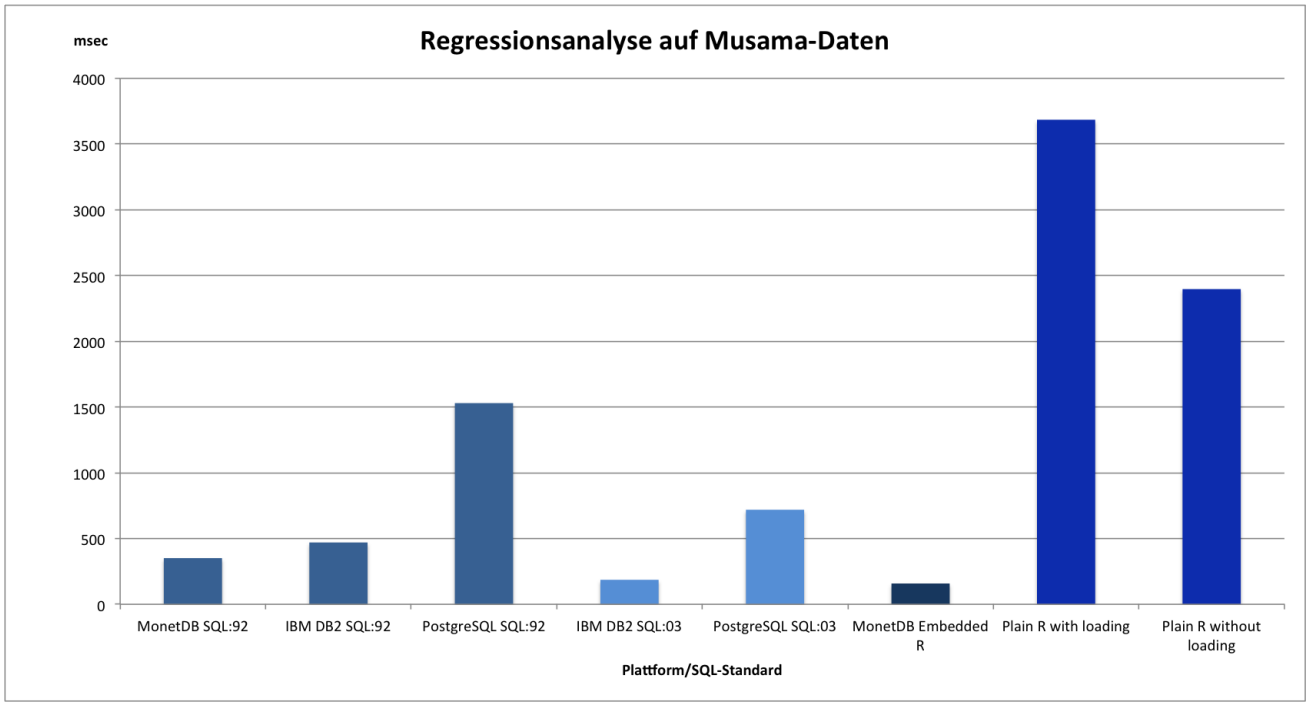

Abbildung 4: Performance-Ergebnisse der eingesetzten Systeme: einfache Regressionsanalyse auf Musama-Daten

9. LITERATUR

[Chi09] Chirkova, R.: Query containment. In: Liu, L.; Özsu, M. T. (Hrsg.): *Encyclopedia of Database Systems*, S. 2249–2253. Springer US, 2009.

[Dit15] Dittrich, J.: The case for small data management. In: Seidl, T.; Ritter, N.; Schöning, H.; Sattler, K.; Härder, T.; Friedrich, S.; Wingerath, W. (Hrsg.): *Datenbanksysteme für Business, Technologie und Web (BTW), 16. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme"(DBIS), 4.-6.3.2015 in Hamburg, Germany. Proceedings*, *LNI*, Band 241, S. 27–28. GI, 2015.

[Est14] Estrin, D.: Small data, where n = me. *Commun. ACM*, Band 57, Nr. 4, S. 32–34, 2014.

[FBH+14] Finger, A.; Bruder, I.; Heuer, A.; Klemkow, M.; Konerow, S.: PageBeat - Zeitreihenanalyse und Datenbanken. In: *Proceedings of the 26th GI-Workshop Grundlagen von Datenbanken, Bozen-Bolzano*, S. 53–58, 2014.

[Gru14] Grunert, H.: Distributed denial of privacy. In: Plödereder, E.; Grunske, L.; Schneider, E.; Ull, D. (Hrsg.): *44. Jahrestagung der Gesellschaft für Informatik, Informatik 2014, Big Data - Komplexität meistern, 22.-26. September 2014 in Stuttgart, Deutschland*, *LNI*, Band 232, S. 2299–2304. GI, 2014.

[Heu15] Heuer, A.: METIS in PArADISE: Provenance Management bei der Auswertung von Sensordatenmengen für die Entwicklung von Assistenzsystemen. In: *Lecture Notes in Informatics, Band 242, BTW 2015 Workshop-Band, 131 – 135*, 2015.

[HKG14] Heuer, A.; Karopka, T.; Geisler, E.: *Einführung in Ambient Assisted Living*. Lehrbuch Projekt BAAL (WeiterBildung in Ambient Assisted Living) der Universität Rostock, 2014.

[HKHT06] Heuer, A.; Kirste, T.; Hoffmann, W.; Timmermann, D.: Coast: Concept for proactive assistive systems and technologies. Bericht, Fakultät für Informatik und Elektrotechnik der Universität Rostock, 2006.

[HMB14] Heuer, A.; Meyer, H.; Bruder, I.: Das digitale Gedächtnis erhalten. *Transfer - Das Steinbeis Magazin*, Jahrgang 2014, Nr. 04, 2014.

[KNY+14] Krüger, F.; Nyolt, M.; Yordanova, K.; Hein, A.; Kirste, T.: Computational State Space Models for Activity and Intention Recognition. A Feasibility Study. *PLOS ONE*, Jahrgang 2014, November 2014. 9(11): e109381. doi:10.1371/journal.pone.0109381.

[Mar15] Markl, V.: Gesprengte Ketten - Smart Data, deklarative Datenanalyse, Apache Flink. *Informatik Spektrum*, Band 38, Nr. 1, S. 10–15, 2015.

[WKL+15] Wilsdorf, P.; Köppl, F.; Lüdtke, S.; Sachse, S.; Svacina, J.; Weu, D.: Abschlusspräsentation Projektgruppe PArADISE 2014/2015, 2015.