FernUniversität in Hagen

– Fakultät für Mathematik und Informatik –

Reproduzierbarkeit in der Datenbanksystem-Forschung

Seminar „Big Data Management“

Lehrgebiet Datenbanken und Informationssysteme

vorgelegt von

Anne Koch

Matrikelnummer: 3217159

Referentin : Prof. Uta Störl

ERKLÄRUNG

Ich erkläre, dass ich die schriftliche Ausarbeitung zum Seminar selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe.

Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht.

Die Versicherung selbstständiger Arbeit gilt auch für enthaltene Zeichnungen, Skizzen oder graphische Darstellungen.

Die Ausarbeitung wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Mit der Abgabe der elektronischen Fassung der endgültigen Version der Ausarbeitung nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate geprüft werden kann und ausschließlich für Prüfungszwecke gespeichert wird.

Dresden, 04. Juli 2022

Anne Koch

1

ABSTRACT

Die Reproduzierbarkeit von Forschungsergebnissen ist ein in allen Wissenschaftsdisziplinen wichtiges und schwieriges Thema. In der Datenbanksystem-Forschung stellt sie durch die stetig zunehmenden Datenmengen eine besondere Herausforderung dar.

Obwohl die wichtigsten Datenbankkonferenzen seit mittlerweile über zehn Jahren Preise für Reproduzierbarkeit vergeben, ist der Anteil reproduzierbarer Forschungsarbeiten weiterhin gering. Um einen Rahmen und Hintergründe für diese aktuelle Situation aufzuzeigen, betrachtet die Arbeit in einem theoretischen Teil den aktuellen Stand der Reproduzierbarkeit wissenschaftlicher Forschung im Bereich Datenbanken, Vor- und Nachteile reproduzierbarer Forschungsarbeiten, Ziele der wissenschaftlichen Gemeinschaft bezüglich Reproduzierbarkeit sowie Gründe für die Nichtreproduzierbarkeit von Studien.

In einem weiteren, praktischen Teil werden Herausforderungen und Lösungsansätze zum Thema Reproduzierbarkeit in der Datenbanksystem-Forschung vorgestellt, die in der Literatur vorgeschlagen werden. Erörtert werden die geeignete Wahl der Forschungsdaten, die Protokollierung der Experimente mittels Workflow- und Provenienzsoftware, die Zusammenfassung der Experimentauswertung durch ausführbare Paper sowie Methoden, die eine langfristige Online-Bereitstellung der Daten und Skripte ermöglichen. Diese Methoden für die Veröffentlichung der Ergebnisse umfassen einen *Digital Object Identifier* zum eindeutigen Auffinden der Ressourcen, die Bereitstellung der Daten entweder in Form eines Datengenerierungsalgorithmus oder der vollständigen Rohdaten inklusive eindeutiger Anweisungen zum Umwandeln dieser in die Eingangsdaten für die Experimente sowie die Bereitstellung der ausführbaren Auswertungsdateien in Form von Jupyter-Notebooks oder Containerimages.

Mit diesen Lösungsansätzen stehen Methoden bereit, die eine Reproduzierbarkeit von Forschungsergebnissen in der Datenbanksystem-Forschung gewährleisten.

INHALTSVERZEICHNIS

1 ABSTRACT iii

2 EINLEITUNG 1
- 2.1 Forschungsfragen . . . 1
- 2.2 Struktur der Arbeit . . . 2

3 GRUNDLAGEN 3
- 3.1 Reproduzierbarkeit . . . 3
- 3.2 Benchmarks . . . 3
- 3.3 Schema-Evolution . . . 4
- 3.4 Datenprovenienz . . . 4

4 AKTUELLER STAND 5
- 4.1 Vor- und Nachteile reproduzierbarer Forschung . . . 5
- 4.2 Ziele bezüglich reproduzierbarer Forschung in der Informatik . 6
- 4.3 Aktuelle Lage der Reproduzierbarkeit wissenschaftlicher Forschung im Bereich Datenbanken . . . 6
- 4.4 Gründe für Nichtreproduzierbarkeit . . . 7

5 HERAUSFORDERUNGEN UND LÖSUNGSANSÄTZE 8
- 5.1 Planungsphase . . . 8
  - 5.1.1 Auswahl und Erstellung der Daten . . . 9
  - 5.1.2 Datenänderungen . . . 9
- 5.2 Simulationsphase . . . 10
  - 5.2.1 Datengenerierungsalgorithmus . . . 10
  - 5.2.2 Workflow- und Provenienzsoftware . . . 10
- 5.3 Auswertungsphase . . . 11
  - 5.3.1 Ausführbare Paper . . . 11
- 5.4 Veröffentlichungsphase . . . 12
  - 5.4.1 Digital Object Identifier . . . 12
  - 5.4.2 Bereitstellung der Daten . . . 13
  - 5.4.3 Containerisierung . . . 15

6 BEISPIEL EINES REPRODUZIERBAREN PAPERS 17
- 6.1 Reproduktion des Papers . . . 18

7 DISKUSSION 19
- 7.1 Nutzung von Benchmarks . . . 19
- 7.2 Langfristige Bereitstellung von Datensätzen . . . 19
- 7.3 Änderungen der Datensätze . . . 19

8 ZUSAMMENFASSUNG UND AUSBLICK 20

A ANHANG 21

LITERATUR 24

2 EINLEITUNG

> Ideally reproducibility should be close to zero effort.
>
> SIGMOD
> Reproducibility-Website
> [ACM22b]

In dieser Seminararbeit wird der aktuelle Stand und die Herausforderungen des Themas Reproduzierbarkeit für die Datenbanksystem-Forschung analysiert.

Wissenschaft basiert auf Austausch und Prüfung von früheren Forschungsergebnissen. [Dat10] beschreiben, dass traditionelle Forschungsarbeiten in Bereichen wie den Naturwissenschaften und Mathematik sowohl den neuartigen Beitrag als auch die Informationen, die erforderlich sind, um diesen zu reproduzieren, wie detaillierte Beschreibungen der verwendeten empirischen Methoden oder mathematische Beweise, enthielten.

Durch die technischen Weiterentwicklungen und die Möglichkeiten des Internet stehen heute jedoch Datenmengen für wissenschaftliche Forschung zur Verfügung, die nicht mehr in der Forschungsarbeit selbst wiedergegeben werden können. Daher müssen andere Wege gefunden werden, um die Reproduzierbarkeit der Ergebnisse zu gewährleisten.

Obwohl der Forschungsgegenstand in der Informatik inhärent deterministischer ist als bspw. in Natur- oder Sozialwissenschaften, wird für die Forschung in der Informatik eine Glaubwürdigkeits- und Reproduzierbarkeitskrise proklamiert [FBS12a], [Chi+13], da viele Forschungsergebnisse nicht reproduzierbar sind. Dies zeigt, dass Reproduzierbarkeit, auch wenn sie wichtig ist, nicht einfach zu erreichen ist.

Damit Forschungsergebnisse in der Datenbanksystem-Forschung nachträglich überprüft werden können, ist es erforderlich, dass die Hardware- und Softwareumgebung sowie die Daten aus der ursprünglichen Forschungsarbeit wiederhergestellt werden können.

2.1 FORSCHUNGSFRAGEN

Diese Arbeit soll dabei insbesondere untersuchen, welchen Einfluss die Auswahl und Erstellung der Datensätze auf die Reproduzierbarkeit haben. Dabei wird auf die folgenden Fragestellungen eingegangen:

- Zur Leistungsmessung von Datenbankmanagementsystemen werden häufig Benchmarks genutzt. Welche Benchmarks existieren und inwie-

fern wirkt sich die Nutzung von Benchmarks auf die Reproduzierbarkeit aus?

- Welche Ansätze gibt es, um (künstlich erzeugte oder reale) Datensätze langfristig bereit zu stellen und für nachträgliche Arbeiten auswertbar zu machen?
- Welchen Einfluss haben Änderungen der Datensätze über die Zeit (möglicherweise Jahre nach der Ursprungsveröffentlichung)?

2.2 STRUKTUR DER ARBEIT

Die Arbeit ist wie folgt strukturiert:

In Kapitel 3 werden einige Begriffe, die für den Bereich der Reproduzierbarkeit von Bedeutung sind, eingeführt.

Kapitel 4 zum aktuellen Stand führt Vor- und Nachteile reproduzierbarer Forschung sowie Ziele reproduzierbarer Forschung in der Informatik an, gibt eine Übersicht über Untersuchungen zum Anteil reproduzierbarer Arbeiten in der Datenbankforschung und führt außerdem Gründe dafür an, warum manche Forschungsergebnisse nicht reproduzierbar sind.

Kapitel 5 widmet sich Herausforderungen der Reproduzierbarkeit in der Datenbankforschung und einigen Lösungsansätzen dafür, die in einzelnen Phasen einer Forschungsarbeit auftreten, von der Planung über die Durchführung der Simulationen, deren Auswertung bis hin zur Veröffentlichung der Ergebnisse.

In Kapitel 6 wird beispielhaft eine reproduzierbare Forschungsarbeit vorgestellt, deren veröffentlichtes Dockerimage ausgeführt und reproduziert wurde. Die erzeugten Messwerte und Abbildungen werden den Originaldaten aus dem publizierten Paper gegenübergestellt.

In Kapitel 7 Diskussion werden die in der Arbeit gefundenen Antworten auf die in Abschnitt 2.1 formulierten Forschungsfragen zusammengefasst und offen gebliebene Fragen erwähnt.

In Kapitel 8 werden die Erkenntnisse aus der Arbeit zusammengefasst und Vorschläge für weiterführende Arbeiten gemacht.

3 GRUNDLAGEN

In diesem Kapitel werden einige für die Reproduzierbarkeit von Forschungsergebnissen wichtige Begriffe definiert, die für das Verständnis des weiteren Textes erforderlich sind.

3.1 REPRODUZIERBARKEIT

[ACM22a] unterscheidet drei Arten der Reproduzierbarkeit danach, durch wen und wie die Forschungsergebnisse erzielt wurden:

- Wiederholbarkeit (*Repeatability*) - durch dieselbe Forschungsgruppe, mit demselben Versuchsaufbau

  Die Messung kann von der selben Forschungsgruppe in mehreren Versuchen mit demselben Messverfahren, demselben Messsystem unter denselben Betriebsbedingungen und am selben Ort ermittelt werden.

- Reproduzierbarkeit (*Reproducibility*) - durch eine andere Forschungsgruppe, mit demselben Versuchsaufbau

  Die Messung kann von einer anderen Forschungsgruppe in mehreren Versuchen mit demselben Messverfahren, demselben Messsystem unter denselben Betriebsbedingungen und am selben Ort ermittelt werden.

- Replizierbarkeit (*Replicability*) - durch eine andere Forschungsgruppe, mit einem anderen Versuchsaufbau

  Die Messung kann von einer anderen Forschungsgruppe in mehreren Versuchen mit einem anderen Messverfahren, einem anderen Messsystem an einem anderen Ort ermittelt werden.

Diese Termini werden jedoch nicht einheitlich verwendet. So schreiben beispielsweise [Bon+11] von *Repeatability*, obwohl es in ihrer Untersuchung um die Reproduzierung der Forschungsergebnisse durch die Reviewer der SIGMOD-Konferenz geht.

3.2 BENCHMARKS

Häufig werden zur Ermittlung von Leistungswerten von Datenbanken Benchmarks eingesetzt. Laut [LÖ18] sind Datenbank-Benchmarks reproduzierbare experimentelle Frameworks zur Charakterisierung und zum Vergleich der Leistung (Zeit, Speicher oder Qualität) eines Datenbanksystems oder von

Algorithmen auf diesen Systemen. In einem solchen experimentellen Framework werden das *System Under Test* (SUT - zu testendes System), die Arbeitslast, Metriken und Experimente definiert.

Das *System Under Test* umfasst ein Datenbanksystem und seine Ausführungsumgebung, die aus Betriebssystemdiensten besteht, die auf Hardwarekomponenten aufbauen. Ein Benchmark kann Beschränkungen für die Ausführungsumgebung festlegen, um sicherzustellen, dass die Leistung verschiedener Datenbanksysteme vergleichbar ist.

Die Arbeitslast besteht in der Regel aus einem synthetischen Datensatz, der entweder explizit als Datei vorgegeben oder durch einen Datengenerierungsalgorithmus definiert wird.

Zur Verfügung stehen Micro-Benchmarks und Standard-Benchmarks. [Man+09] beschreiben Micro-Benchmarks als Spezialsoftware, die bestimmte Teile eines größeren Systems isolieren, z.B. einzelne Datenbank-Operatoren (select, join usw.), wohingegen Standard-Benchmarks echte Praxis-Szenarien simulieren. Ein häufig genutzter Benchmark ist beispielsweise der TPC-H zur Entscheidungsfindung des *Transaction Processing Performance Council* (TPC).

3.3 SCHEMA-EVOLUTION

Gemäß [LÖ18] ergibt sich die Schema-Evolution aus der Notwendigkeit aktuelle Datenbestände zu erhalten, wenn Änderungen am Datenbank-Schema vorgenommen werden.

Da es im Lebenszyklus von Datenbanken häufig zu Schema-Änderungen kommt, kommt der Schema-Evolution große Bedeutung zu und sie ist auch Thema intensiver Forschung.

3.4 DATENPROVENIENZ

Gemäß [LÖ18] wird als Datenprovenienz die Protokollierung der Herkunft von Daten (in einer Datenbank, einem Dokument oder einem Repository) einschließlich einer Begründung für die Aufnahme der Daten bezeichnet.

Die Protokollierung der Herkunft von Daten von Experimenten durch Aufzeichnung verwendeter Bedingungen und Parameter ist insbesondere deshalb wichtig, um Experimente nachträglich nachvollziehen und reproduzieren zu können.

4 AKTUELLER STAND

In diesem Kapitel werden Vor- und Nachteile reproduzierbarer Forschungsarbeiten, Ziele der wissenschaftlichen Gemeinschaft bezüglich Reproduzierbarkeit, die aktuelle Lage der Reproduzierbarkeit wissenschaftlicher Forschung im Bereich Datenbanken sowie Gründe für die Nichtreproduzierbarkeit von Studien angeführt.

4.1 VOR- UND NACHTEILE REPRODUZIERBARER FORSCHUNG

Die große Mehrheit der untersuchten Arbeiten führen in der einen oder anderen Weise an, dass die Reproduzierbarkeit wissenschaftlicher Arbeit einen Grundbestandteil guter wissenschaftlicher Praxis darstellt.

Als Vorteile bereitgestellter Daten und reproduzierbarer Studien führen [Ten+11], [Chi+13] und [FBS12a] an:

- Möglichkeit der Reproduzierung, Verifizierung und des Vergleichs der Studien mit gleicher Konfiguration
- Möglichkeit anderer Auswertungen, Nutzung geänderter Parameter, Erweiterung der Experimente
- Gute Verwaltung und langfristige Bereitstellung der Daten
- Vermeidung wiederholter Datenerstellung und damit Ressourcenoptimierung
- Portierbarkeit reproduzierbarer Software
- Verhinderung von wissenschaftlichem Fehlverhalten durch Datenfälschung
- Möglichkeit der Nutzung als Trainingsmaterial für nachfolgende Forschende
- Verstärkte Zitierung und Sichtbarkeit von Papern, die reproduzierbare Experimente enthalten

Allerdings gibt es jedoch auch vereinzelte Stimmen, die den Bemühungen um reproduzierbare Forschung kritisch gegenüber stehen. [Dru12] kritisiert die Betonung der Reproduzierbarkeit als grundlegenden Bestandteil wissenschaftlicher Arbeit sowie die Vorstellung, dass es genau eine korrekte wissenschaftliche Methode gäbe, in der Forschungsarbeiten inkrementell aufeinander aufbauen. Außerdem befürchtet er, dass die Forderung zur Veröffentlichung von Daten und Code zu Misstrauen unter Forschenden, übermäßiger Zusatzarbeit für Reviewer und Annahme von Papern aufgrund enger technischer Kriterien führen kann. Schließlich führt er an, dass es schon

immer Fehlverhalten in der Wissenschaft gegeben hat, was sich jedoch wenig auf den Fortschritt der Forschung insgesamt ausgewirkt hat, und kommt insgesamt zu dem Schluss, dass die Entscheidung über die Veröffentlichung von Code und Daten besser den einzelnen Forschenden zu überlassen sei und diesbezügliche Vorgaben durch Fachzeitschriften oder Geldgebende der wissenschaftlichen Gemeinschaft nicht dienen.

4.2 ZIELE BEZÜGLICH REPRODUZIERBARER FORSCHUNG IN DER INFORMATIK

2009 trafen sich Wissenschaftlerinnen, Anwälte, Herausgebende von Fachzeitschriften und Förderer zu einem Runden Tisch an der Yale Law School, um Möglichkeiten zu erörtern, wie Daten und Code in traditionelle Forschungspublikationen im Bereich Informatik integriert werden können [Dat10]. Dabei erarbeiteten sie Empfehlungen für Forschende, Geldgebende und Herausgebende von Fachzeitschriften, die vom Beispiel der Genomforschungsgemeinschaft inspiriert war, die 1996 mit den Bermuda-Prinzipien ein kooperative Strategie ausgearbeitet hatte, die die Praxis des Datenaustauschs in dieser Forschungsgemeinschaft geprägt hat. Die vom Runden Tisch veröffentlichten Empfehlungen für Forschende umfassten die folgenden Punkte:

- Bereitstellung der Quellcodeversion und der Daten, die für die Berechnung der Ergebnisse genutzt wurden;
- Vergabe einer einmaligen ID für jede Version des veröffentlichten Codes und Aktualisierung derselben bei jeder Code- oder Datenänderung;
- Beschreibung der für die Veröffentlichung verwendeten Hardwareumgebung und der Softwareversion inkl. stabiler Links zu dem entsprechenden Code und den Daten, ggf. auch einer virtuellen Maschine;
- Nutzung offener Lizenzen für Code;
- Nutzung von Open-Access-Verträgen für veröffentlichte Arbeiten und Bereitstellung von Preprints auf Plattformen wie arXiv.org, PubMed Central oder Harvard's Dataverse Network;
- Veröffentlichung der Daten und des Codes in nicht-proprietären Formaten, die möglichst noch langfristig lesbar sein werden.

4.3 AKTUELLE LAGE DER REPRODUZIERBARKEIT WISSENSCHAFTLICHER FORSCHUNG IM BEREICH DATENBANKEN

Auf der jährlichen Konferenz der Fachgruppe der Association for Computing Machinery (ACM) für Datenmanagement SIGMOD wurden von 2008

bis 2012 experimentelle Reproduzierbarkeits- und Wiederholbarkeitsanstrengungen unternommen und seit 2015 werden Preise für die Reproduzierbarkeit der vorgestellten Paper vergeben [ACM22c]. Ebenso werden bei der Konferenz zu Datenbanken für die Verarbeitung sehr großer Datenmengen (VLDB - Very Large Data Bases) Bemühungen zur Förderung der Reproduzierbarkeit unternommen. Trotz dieser Anstrengungen hat jedoch laut [Paw+19] in dem untersuchten Zeitraum von 2015-2019 die Anzahl der bei den entsprechenden Konferenzen veröffentlichten reproduzierbaren Paper nicht zugenommen und lag im Durchschnitt bei 8 %.

Während der in mehr als zehn Jahre gesammelten Erfahrungen mit Reproduzierbarkeitsbemühungen durch die Konferenzen wurden die Bedingungen und Anweisungen für die Bereitstellung reproduzierbarer Paper immer weiter angepasst und verfeinert.

4.4 GRÜNDE FÜR NICHTREPRODUZIERBARKEIT

In [Ten+11] wurde eine Umfrage zu aktuellen Gewohnheiten bezüglich der Bereitstellung von Forschungsdaten und Ansichten zu Hindernissen und Förderungsmöglichkeiten durchgeführt, die von 1329 Forschenden verschiedenster Disziplinen beantwortet wurde, von denen 9% aus den Fachbereichen Informatik und Ingenieurwissenschaften stammen. Von diesen Personen wurden folgende Gründe angeführt, warum sie ihre Daten nicht elektronisch zur Verfügung stellen:

- Fehlende Zeit: 53,6%
- Fehlende Finanzierung: 39,6%
- Fehlende Rechte für Veröffentlichung der Daten: 24,1%
- Fehlender Speicherort für die Daten: 23,5%
- Fehlende Standards: 19,8%
- Fehlende Forderung durch die geldgebende Institution: 17,4%
- Daten werden nicht gebraucht: 15,0%
- Sonstige Gründe: 14,6%
- Daten sollten nicht bereitgestellt werden: 14,4%

Spezifisch im Bereich Datenbanken führt [Bon+11] als Gründe für die Nichtteilnahme an den Reproduzierbarkeitstests der SIGMOD-Konferenzen Rechte am geistigen Eigentum der Software, vertrauliche Daten und besondere Hardwareanforderungen an.

5 HERAUSFORDERUNGEN UND LÖSUNGSANSÄTZE

Dieses Kapitel behandelt Herausforderungen, die bei reproduzierbarer Forschung auftreten, sowie Lösungsansätze, wie damit umgegangen werden kann. Das Kapitel umfasst einige grundlegende Überlegungen zur Planung von Forschungsarbeiten in Bezug auf Reproduzierbarkeit sowie Abschnitte zu den Herausforderungen in den einzelnen Phasen der Erstellung einer Forschungsarbeit.

Für Reproduzierbarkeit kann in zwei verschiedenen Weisen gesorgt werden: erstens durch Einplanen der dafür nötigen Schritte von Anfang an und zweitens durch nachträgliche Bereitstellung der Experimentdaten nach Abschluss der Arbeit. Wenn die Bereitstellung nicht von Beginn der Forschung an eingeplant war, ist es erforderlich, nachträglich eine Beschreibung zu erstellen, wie die Experimentergebnisse erzielt wurden.

Ist die Reproduzierbarkeit geplant, gibt es Best Practises dafür, die in den nachfolgenden Abschnitten beschrieben werden.

[Chi+13] unterteilen eine Forschungsarbeit in drei Phasen:

SIMULATIONSPHASE: Zuerst werden umfangreiche Simulationen vorbereitet und durchgeführt, aus denen sich die rohen Simulationsdaten ergeben. Diese Phase ist üblicherweise zeitaufwändig und lässt sich nicht leicht reproduzieren. Daher werden die Rohdaten zusammen mit allen Schritten, die für eine Reproduktion der Daten erforderlich sind, archiviert.

AUSWERTUNGSPHASE: Als nächstes werden die Daten analysiert und ausgewertet. Beispielsweise werden die Daten in Diagrammen zusammengefasst, aus denen sich die Qualität der abgeleiteten Ergebnisse ablesen lässt.

VERÖFFENTLICHUNGSPHASE: Abschließend werden sowohl die Grafiken als auch andere Ergebnisse in die wissenschaftliche Arbeit eingefügt.

Zur Strukturierung werden die Herausforderungen in den nachfolgenden Abschnitten in eine chronologische Abfolge von Planung und diesen drei Phasen eingeteilt, um darzustellen, welche Überlegungen in den einzelnen Phasen jeweils zu treffen sind.

5.1 PLANUNGSPHASE

In der Planungsphase müssen grundsätzliche Entscheidungen für die Forschungsarbeit getroffen werden.

Da in der Datenbanksystem-Forschung häufig große Datenmengen genutzt werden, kommt Überlegungen, wie diese erstellt, gespeichert, archiviert und anderen Nutzenden zur Verfügung gestellt werden können, große

Bedeutung zu. Gegebenenfalls müssen Maßnahmen getroffen werden für den Fall, dass es zu Änderungen am Datenbestand kommt. Auf diesbezügliche Überlegungen wird in den nachfolgenden Abschnitten ausführlicher eingegangen.

Insbesondere, wenn es sich bei der Forschungsarbeit um eine Arbeit handelt, die die Leistung von Datenbank-Managementsystemen misst, ist es von Bedeutung, die verwendete Hardware genau zu dokumentieren. Denn im Gegensatz zur verwendeten Software, lässt sich die Hardware nicht containerisieren und als Gesamtpaket für nachträgliche Reproduktionen zur Verfügung stellen.

Da die in einem Forschungsinstitut verwendete Hardware sich im Allgemeinen nicht ständig und in schneller Abfolge ändert, ist es, wie [GB12] argumentieren, möglich, ein Verzeichnis der vorhandenen Hardware zu führen und bei Änderungen zu aktualisieren. Dann lassen sich zu jeder Publikation die Daten der aktuellen Hardware schnell zusammenstellen.

5.1.1 *Auswahl und Erstellung der Daten*

Bei der Planung des Experiments ist zu entscheiden, ob als Datengrundlage künstlich erzeugte Daten oder echte Daten verwendet werden sollen. Bei der Verwendung synthetischer Daten bieten sich häufig Benchmarks an, die für viele verschiedene Einsatzzwecke zur Verfügung stehen.

[Man+09] stellen Kriterien zur Datenauswahl für Datenbankforschungsarbeiten vor, die die Leistung der untersuchten Systeme messen. Zur Auswahl stehen Micro-Benchmarks, Standard-Benchmarks und echte Anwendungen. Bezüglich der Reproduzierbarkeit haben Micro-Benchmarks den Vorteil, dass die Datenmenge skalierbar ist und sie einfach einzurichten und auszuführen sind, jedoch den Nachteil, dass die Metriken nicht standardisiert und die Benchmarks schlecht vergleichbar sind. Der Vorteil von Standard-Benchmarks ist, dass sie öffentlich verfügbar sind und skalierbare Datensätze und Arbeitslasten zur Verfügung stehen. Nachteile von Standard-Benchmarks sind, dass sie oft überholt sind, da die Standardisierung lange dauert und dass sie häufig sehr groß und umständlich auszuführen sind. Bspw. beträgt die Datenbankgröße des TPC-H-Benchmarks zwischen 1 GB und 100 000 GB.

In Bezug auf die Verwendung echter Daten führen [BMS20] an, dass es früher schwierig war, Zugriff auf echte Datensätze zu erhalten, weshalb frühere Studien hauptsächlich an nicht öffentlich zugänglichen Daten durchgeführt wurden. Durch die Ausbreitung von Open-Source-Software und -Daten und den Zugang zu Code-Repositorys ergeben sich jedoch ganz neue Forschungsmöglichkeiten, zum Beispiel anhand von GitHub und Wikipedia.

5.1.2 *Datenänderungen*

Bestehen Datensätze über eine lange Zeit, kommt es in vielen Fällen zu Änderungen am Datenbestand, sowohl was die Einträge betrifft als auch durch

Schema-Evolution. Beides muss berücksichtigt werden, wenn die Rohdaten nicht als vollständiger Datensatz, sondern beispielsweise als Link bereitgestellt werden sollen. [Paw+19] argumentieren jedoch dagegen, die Experimentdaten nur in Form eines Links bereitzustellen, denn wenn sich die Daten ändern, ist es nicht möglich, aus den Rohdaten exakt die Eingangsdaten wie in der originalen Arbeit zu gewinnen. Werden hingegen nur die aus den Rohdaten gewonnen Eingangsdaten zur Verfügung gestellt, wird dadurch der Vorgang der Datenaufbereitung und die Details der gewonnenen Informationen verdeckt und es bleibt unklar, wie die Rohdaten in die Eingangsdaten umgewandelt wurden.

Je nach Forschungsgegenstand kann die Antwort auf die Frage, ob die reproduzierte Forschung anhand von identischen oder nur äquivalenten Daten ausgeführt werden sollte, unterschiedlich ausfallen. Ggf. reichen äquivalente Daten aus, um zu zeigen, dass das reproduzierte Experiment gleichwertige Ergebnisse erzielt. Allerdings kann es schwierig sein, nachzuweisen, dass der Unterschied in den Daten nicht so groß ist, dass er selbst Auswirkungen auf das Ergebnis hat und dieses verfälscht. Diese Gefahr wird vermieden, wenn tatsächlich mit identischen Daten gearbeitet wird.

5.2 SIMULATIONSPHASE

In der Simulationsphase können durch den Einsatz von Datengenerierungsalgorithmen sowie Workflow- und Provenienzsoftware die Grundlagen für eine spätere Reproduzierbarkeit der Forschungsergebnisse gelegt werden.

5.2.1 *Datengenerierungsalgorithmus*

Wenn als Datengrundlage kein bestehender Datenbestand verwendet werden soll, sondern Daten für die Arbeit neu generiert werden, kann dafür ein Datengenerierungsalgorithmus genutzt werden. Derartige Algorithmen erzeugen synthetische Datensätze und existieren für eine Vielzahl von Anwendungsbereichen.

Für die reproduzierbare Forschung im Bereich Datenbanksysteme sind sie vor allem daher interessant, da sie es ermöglichen, umfangreiche Rohdatenbestände zu erstellen, bei denen es unpraktikabel wäre, die Rohdaten selbst reproduzierbar zu archivieren.

Beispielsweise stehen mit DBGEN und QGEN [Phi22] Programme zur Verfügung, um Datenbankeinträge sowie Querys für den TPC-H-Benchmark zu erstellen.

5.2.2 *Workflow- und Provenienzsoftware*

Mit Hilfe von Workflows lassen sich Experimente automatisieren und gleichzeitig dokumentieren. Provenienzsoftware verwaltet die Ressourcen und Parameter, die zur Erstellung von Experimenten und Diagrammen verwendet werden. Beim Einsatz von Workflow- und Provenienzsoftware werden die

für jedes Experiment verwendeten Programme, Skripte und Parameter protokolliert und mit den erzeugten Daten und Grafiken verlinkt, so dass immer ermittelt werden kann, welche Bedingungen zu welchem Ergebnis geführt haben. Die automatisierte und möglichst vollständige Aufzeichnung dieser Informationen ist sowohl während der Forschungstätigkeit nützlich, um interessante Wertekombinationen zu finden, deren weitergehende Untersuchung lohnt, als auch danach bei einer Reproduktion der Ergebnisse, um die Einstellungen, die zu den veröffentlichten Ergebnissen geführt haben, zu ermitteln. Als Beispiele für Workflow- und Provenienzsoftware führen [FBS12b] an:

- das Workflow-Managementsystem *Pegasus*: `https://pegasus.isi.edu`
- die Softwareanwendung zur Analyse und Modellierung wissenschaftlicher Daten *Kepler*: `https://kepler-project.org`
- das Managementsystem für wissenschaftliche Workflows und Provenienz *VisTrails*: `http://www.vistrails.org` (seit 2018 nicht mehr weiterentwickelt)
- das Softwarepaket für multidimensionale Datenanalyse und reproduzierbare Computerexperimente *Madagaskar*: `https://www.reproducibility.org`

5.3 AUSWERTUNGSPHASE

In der Auswertungsphase werden aus den gewonnen Rohdaten die Forschungsergebnisse ermittelt. Sollen diese reproduzierbar sein, müssen die dafür genutzten Verfahren so aufbereitet werden, dass sie veröffentlicht werden können. Dies kann zum Beispiel durch ausführbare Paper geschehen.

5.3.1 *Ausführbare Paper*

Ausführbare Paper sind Studien, die mit Links zu ausführbarem Code versehen oder in Form von beispielsweise Jupyter-Notebooks bereitgestellt werden und es dadurch ermöglichen, das Zustandekommen der Experimentergebnisse und Grafiken nachzuvollziehen.

Je nach verwendetem System sind die Ergebnisse transparenter oder eher Blackbox-artig. [FBS12b] haben die folgenden Kriterien ermittelt, anhand derer sich Experimente bezüglich ihres Umfangs oder ihrer Reproduzierbarkeit charakterisieren lassen:

- *Tiefe*: der Detaillierungsgrad der bereitgestellten oder archivierten Experimente:
  - Wertetabellen, die zu einer Arbeit bereitgestellt werden (der Standard heutzutage);

  - das Skript (oder die Tabellenkalkulationsdatei), das für die Berechnung der in dem Paper enthaltenen Abbildungen verwendet wurde, sowie die entsprechenden Wertetabellen;
  - die während der Experimente gemessenen Rohdaten sowie die Skripte, die verwendet wurden, um daraus verschiedene Datensätze abzuleiten;
  - die Experimentdaten (Systemkonfiguration und -initialisierung, Skripte, Arbeitslasten, Messungsprotokolle), die für die Erstellung der Rohdaten verwendet wurden;
  - das Softwaresystem als White-Box (Quelldatei, Konfigurationsdateien, Build-Umgebung) oder Black-Box (ausführbare Datei), in der die Experimente ausgeführt wurden.

- *Portabilität*: die Umgebung, in der die Ergebnisse reproduziert werden können:
  - die Originalumgebung (die Person, die die Experimente durchgeführt hat, kann sie auf ihrem eigenen Rechner wiederholen.
  - eine ähnliche Umgebung (dasselbe Betriebssystem aber auf einem anderen Rechner) oder
  - eine andere Umgebung (ein anderes Betriebssystem oder ein anderer Rechner).

- *Abdeckung*: Anteil der reproduzierbaren Experimente:
  - teilweise Reproduzierbarkeit oder
  - vollständige Reproduzierbarkeit.

5.4 VERÖFFENTLICHUNGSPHASE

Nach Abschluss der eigentlichen Forschungsarbeit und der Auswertung der Ergebnisse wird die Arbeit veröffentlicht. Dazu sind Überlegungen erforderlich, wie sichergestellt werden kann, dass die veröffentlichten Daten dauerhaft zugänglich sind, welche Daten bereitgestellt werden und in welcher Form die Daten bereitgestellt werden.

5.4.1 *Digital Object Identifier*

Um reproduzierbar zu sein, müssen für ein Paper sowohl die Experimentdaten als auch Code und Konfigurationsanweisungen bereitgestellt werden.

Da sich sowohl Software, Plattformen als auch Anbieter ändern, muss gewährleistet werden, dass die Pakete und Daten auch Jahre nach der Ursprungsveröffentlichung noch auffindbar sind. Dazu dient beispielsweise der *Digital Object Identifier* (DOI - Digitaler Objektbezeichner) [ISO22], ein normiertes Identifikationssystem zur Erstellung, Registrierung und Verwaltung von DOI-Namen.

Obwohl die DOI-Norm keine Vorgaben über die Verwendung und Umsetzung der Namensauflösung macht, ermöglicht das System, Ressourcen wiederzufinden, auch wenn sich in der Zwischenzeit die URL oder der Speicherort geändert haben.

5.4.2 *Bereitstellung der Daten*

Die Daten werden entweder als Datensatz oder, wenn dies unpraktisch ist (z.B. weil die Datenmengen zu groß sind), in Form von Anweisungen zur Erzeugung der Daten bereitgestellt.

[Paw+19] argumentieren, dass die Bereitstellung der Rohdaten nicht ausreicht, um Forschungsergebnisse reproduzierbar zu machen. Als Beispiel führen Sie zur Motivierung ihrer Vorschläge die DBLP-Datenbank [DBL22] an, in der bibliographische Daten im XML-Format gespeichert werden und die aufgrund ihrer Verfügbarkeit und intuitiven Nutzbarkeit häufig als Quelle für echte Daten genutzt wird.

In dem Beispiel werden die in den XML-Einträgen codierten hierarchischen Abhängigkeiten zwischen den Daten für Experimente in Baumstrukturen konvertiert. Es existieren jedoch viele verschiedene Möglichkeiten, um aus XML eine Baumstruktur zu erstellen, abhängig davon, ob Formatierungs- und Zusatzinformationen mit übernommen werden oder nicht. So werden bei der Konvertierung A in Abbildung 5.1 alle XML-Elemente des Beispiels aus Listing 5.1. einschließlich Tag-Namen, Attributschlüsseln und -werten erhalten, wohingegen in Konvertierung B (Abbildung 5.2) alle Attribute und HTML-Tags entfernt werden.

Listing 5.1: Gekürztes Beispiel-XML-Element aus [Paw+19]

```
<article mdate="2017–01–11" key="journals/tods/AugstenBG10">
    <author>Nikolaus Augsten</author>
    <author>Michael H.B&ouml;hlen</author>
    <title>
        The
        <i>pq</i>
        -gram distance between ordered labeled trees.
    </title>
    <year>2010</year>
    <volume>35</volume>
    <journal>ACM Trans. Database Syst.</journal>
    <number>1</number>
    <ee>http://doi.acm.org/10.1145/1670243.1670247</ee>
    <url>db/journals/tods/tods35.html#AugstenBG10</url>
    <pages>4:1-4:36</pages>
</article>
```

Beide Konvertierungsarten wurden in Studien zu Ähnlichkeitsjoins von Bäumen genutzt, die aufgrund der unterschiedlichen Baumstrukturen zu erheblichen Unterschieden bei der Anzahl von Labels in einem Datenbestand, der Anzahl ähnlicher Baumpaare, dem resultierenden Datensatz und den entsprechenden Laufzeiten der Joins führten. Somit ist ersichtlich, dass

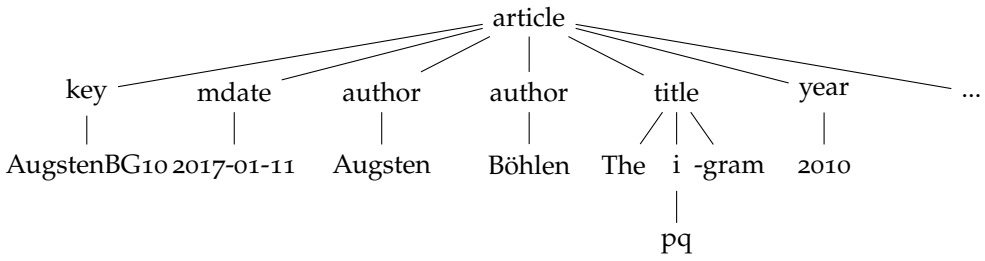


Abbildung 5.1: XML-Baum mit Konvertierung A

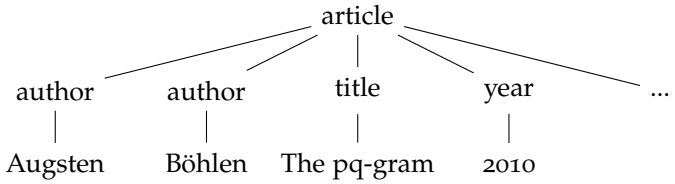


Abbildung 5.2: XML-Baum mit Konvertierung B

die Ergebnisse von Forschungsarbeiten nicht vergleichbar sind, wenn unterschiedliche Konvertierungsstrategien zum Einsatz kommen. Jedoch werden selten genaue Angaben zur Aufbereitung der Rohdaten bereitgestellt. Daher schlagen die Autoren das RPI-Reproduzierbarkeitsmodell vor, mit dem Daten reproduzierbar gemacht werden können.

5.4.2.1 *RPI-Datenreproduzierbarkeitsmodell*

Das Modell besteht aus drei Elementen:

- Rohdaten (R - *raw data*): ein Schnappschuss des Datensatzes im Originalformat, aus dem die Versuchsdaten erstellt werden. Üblicherweise werden Rohdaten weiterverarbeitet und in das gewünschte Datenformat gebracht.

  Es genügt nicht, einen Link zu der Datenquelle bereitzustellen, da dieser Link ungültig werden kann oder sich die unter dem Link bereitgestellten Daten weiterentwickeln können und dann zu anderen Ergebnissen führen können.

- Verarbeitungsanweisungen (P - *preparation instructions*): vollständige und unmissverständliche Anweisungen zur Verarbeitung der Rohdaten in die Eingangsdaten. Diese können in Textform (im Rahmen einer Veröffentlichung) oder idealerweise als automatisiertes Script vorliegen.

- Eingangsdaten (I - *input data*): eine bezüglich Inhalt und Format exakte Kopie der Daten, wie sie für das Experiment verwendet wurden. Es sind keine weiteren Bearbeitungsschritte erforderlich, um die Daten für das Experiment zu nutzen.

  Um eine Verifizierung der Daten zu ermöglichen, kann eine Checksumme zu den Daten angegeben werden.

Werden synthetische Eingangsdaten genutzt, insbesondere bei Zufallswerten, muss sichergestellt werden, dass das Erzeugungsverfahren deterministisch abläuft.

5.4.3 *Containerisierung*

Hardware, Software und das Internet sind kurzlebige Medien. Forschung hingegen hat langfristige Auswirkungen und Studien werden häufig auch Jahre nach ihrer Veröffentlichung zitiert. Die Aufgabe, die Ressourcen für eine Studie langfristig verfügbar zu machen, ist daher keine einfache. Bei der Erstellung dieser Arbeit konnte beispielsweise bereits auf nur einige Jahre alte Quellen, die sich explizit dem Thema Reproduzierbarkeit widmeten, nicht mehr zugegriffen werden. Dies betraf beispielsweise das von [FBS12a] zitierte System für wissenschaftliche Workflows und Provenienzverwaltung *VisTrails*, dessen Entwicklung 2018 eingestellt wurde, sowie sogar die Reproduzierbarkeits-Website der VLDB-Konferenz `https://vldb-repro.com/`, deren Sicherheitszertifikat derzeit nicht aktuell ist.

Es müssen daher Wege gefunden werden, um Quellen möglichst langfristig und sicher zugänglich zu machen.

[MS21] stellen in einem Tutorial Lösungsansätze vor, die dazu beitragen sollen, Forschungsergebnisse langfristig über Jahrzehnte reproduzierbar zu machen.

Als Herausforderungen bei der Bereitstellung von Forschungsergebnissen führen sie an, dass Reproduktionspakete, die auf externe Software zugreifen, unbenutzbar werden können. Als Gründe dafür nennen sie:

- Änderungen der Details des Basissystems
- Änderungen der Konfiguration
- Änderungen des Speicherorts von Repositorys
- Projektumzug auf andere Hosts
- Nichtverfügbarkeit externer Software
- Verschwinden von Download-Links

Aus diesem Grunde schlagen sie vor, sämtliche zur Reproduktion der Ergebnisse erforderlichen Artefakte einschließlich aller erforderlicher Software in einem Docker-Container zu speichern, der heruntergeladen und ohne Internetverbindung ausgeführt werden kann. Ein solches Reproduktionspaket umfasst:

- Generierungscode zum Erzeugen der Daten oder bereits bestehende Eingangsdaten
- Verarbeitungs- und Messungscode
- die erzeugten oder abgeleiteten Daten und

- Visualisierungsskripte, mit denen Grafiken und Diagramme erstellt werden können.

Das gesamte Paket ist gut zu kommentieren, damit auch Personen, die nicht rückfragen können oder wollen, die Ergebnisse reproduzieren können.

Vor dem Veröffentlichen des Publikationsrepositorys sollten Commits bereinigt werden, da für diesen Anwendungsfall nicht die vollständige Entwicklungs- und Versionierungsgeschichte des Repositorys von Interesse ist, sondern nur die nutzbaren Ergebnisse.

Bei der Auswahl der Software ist Open-Source-Technologien der Vorzug zu geben, da diese in großem Maße in Forschung in der Industrie eingesetzt werden, die ein großes Interesse (und die Mittel) hat, diese Software zu warten.

Falls die Nutzung proprietärer Lizenzen unumgänglich ist, muss genau dokumentiert werden, welche Version zum Einsatz kam, damit nachfolgende Forschende bei Bedarf dieselbe Version beschaffen können, und alle erforderlichen Informationen erhalten, um ein kommerzielles DBMS korrekt konfigurieren zu können.

Falls es auch nicht möglich ist, Zugriff auf alle Build-Dateien zu gewähren, sollten zumindest die Ergebnisdaten bereitgestellt werden, damit die Nachbearbeitung der Daten nachvollzogen werden kann.

6

BEISPIEL EINES REPRODUZIERBAREN PAPERS

In diesem Kapitel wird als Beispiel für ein reproduzierbares Paper eines der Siegerpaper des SIGMOD Reproducibility-Awards vorgestellt und reproduziert: *Functional-Style SQL UDFs With a Capital 'F'* (Funktionale benutzerdefinierte SQL-Funktionen mit großem F) [DG20] von C. Duta und T. Grust (Abbildung 6.1). Die Forschungsarbeit behandelt die automatische Optimierung benutzerdefinierter Funktionen in SQL.

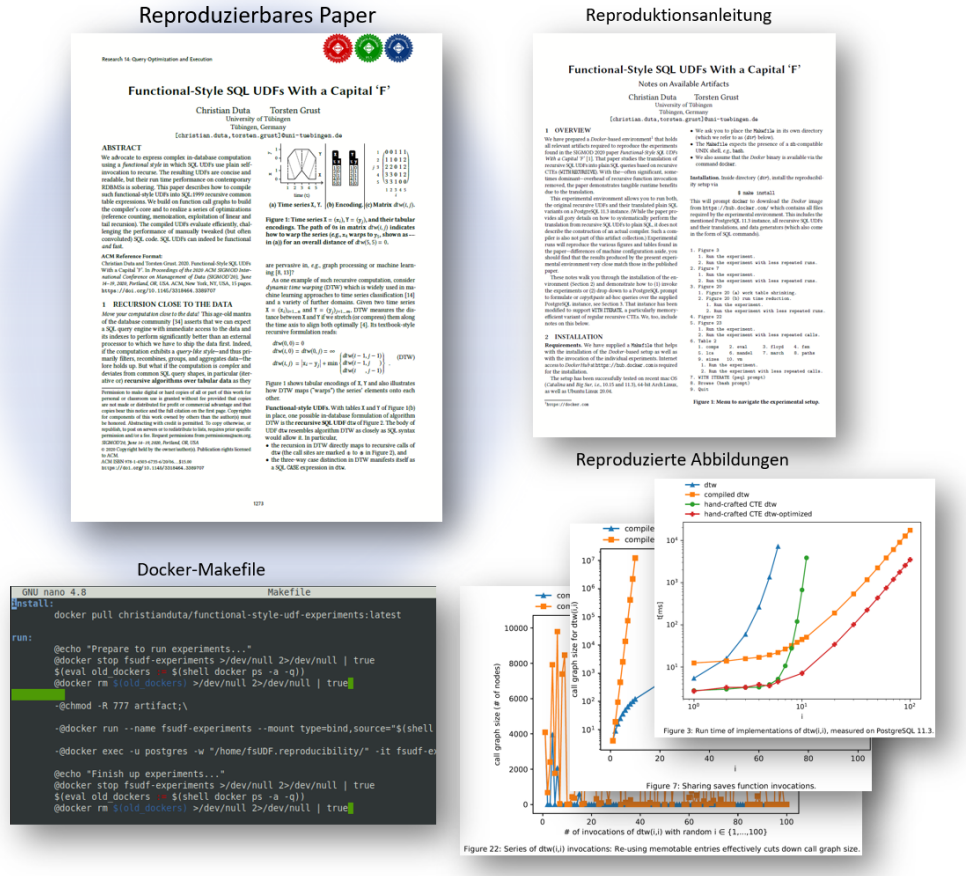


Abbildung 6.1: Ein reproduzierbares Paper [DG20]

Laut dem Reviewer der Konferenz zeichnet sich das Paper wie folgt aus: Die Reproduzierbarkeit dieser Arbeit ist beispielhaft. Alle Softwarekomponenten sind in einem plattformunabhängigen Dockerimage enthalten. Über einen einzigen Kommandozeilenaufruf lassen sich alle Diagramme und Tabellen erneut erstellen. Die bei der Reproduktion gemessenen Absolutwerte entsprechen den in dem Paper angegeben Werten in ausreichendem Maße [Bog21] (Übersetzung der Autorin).

Die per DOI auffindbaren Onlineressourcen umfassen das eigentliche Paper, eine Anleitung zur Reproduktion der Experimente und eine ZIP-Datei mit dem Makefile zur Installation des Dockerimages sowie zum Durchführen der Experimente. Die Anleitung ist gut verständlich.

6.1 REPRODUKTION DES PAPERS

Zur Reproduktion der Berechnungen wurde das Dockerimage auf einem Laptop nicht mehr ganz neuen Datums installiert (Konfiguration siehe Tabelle 6.1).

	Paper	Reproduktion
Plattform	64-bit Linux x86	64-bit Linux x86, Docker 3.1.0
Gerät	Unbekannt	Dell Vostro Laptop (2011)
CPU	Intel Core i7	Intel Core i5
Prozessoren	8	4
GHz	3,66	1,33
RAM	64 GB	4 GB

Tabelle 6.1: Hardware- und Softwarekonfiguration

Wie auch im Reviewbericht beschrieben, konnten die Messergebnisse reproduziert werden. Die erstellten Abbildungen ähneln denen aus dem Originalpaper sowohl in Form als auch bezüglich der Werte, siehe den Vergleich in Abbildung 6.2 (weitere reproduzierte Abbildungen im Anhang).

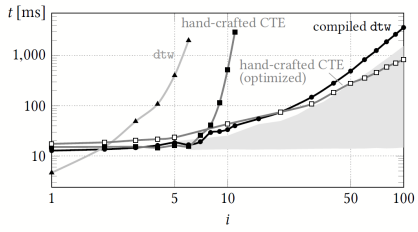


Figure 3: Run time of implementations of `dtw(i,i)`, measured on PostgreSQL 11.3 (logarithmic scales). This work aims for the grey area of low run times.

(a)

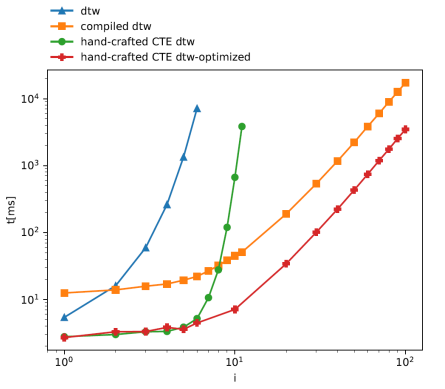


Figure 3: Run time of implementations of dtw(i,i), measured on PostgreSQL 11.3.

(b)

Abbildung 6.2: Vergleich von Abbildung 3 des Originalpapers (a) und des Reproduktionsergebnisses (b)

Die für Tabelle 2 ermittelten absoluten Laufzeiten sind länger als im Originalpaper angegeben (siehe Anhang), was sich jedoch auf die leistungsschwächere Hardware zurückführen lässt. Wie auch im Reviewbericht angegeben, waren die durch die Optimierungen erzielten Verbesserungen in den Reproduktionsberechnungen sogar besser als im Originalpaper.

7 DISKUSSION

Dieses Kapitel erörtert die in Kapitel 5 vorgestellten Lösungsansätze in Bezug auf die in der Einleitung 2.1 gestellten Forschungsfragen.

7.1 NUTZUNG VON BENCHMARKS

Wie in den Abschnitten 5.1.1 und 5.2.1 erörtert, eignen sich (zumindest Standard-)Benchmarks hinsichtlich Reproduktion gut als Datengrundlage, da sie öffentlich verfügbar sind und Datengenerierungsalgorithmen dafür zur Verfügung stehen. Daher reicht es in Forschungsarbeiten aus, Zugriff auf die korrekte Version dieses Algorithmus' bereit zu stellen.

Der Vorteil des Einsatzes von Datengenerierungsalgorithmen ist, dass nicht die teilweise erheblich großen Datensätze archiviert werden müssen.

Genauere Informationen zu den Auswirkungen von Benchmark-Datensätzen auf die Reproduzierbarkeit konnten der Literatur allerdings nicht entnommen werden.

7.2 LANGFRISTIGE BEREITSTELLUNG VON DATENSÄTZEN

Die Frage nach Möglichkeiten einer langfristigen Bereitstellung von Datensätzen wurde in den untersuchten Arbeiten ausführlich beantwortet. Mit dem RPI-Datenreproduzierbarkeitsmodell (Abschnitt 5.4.2.1, Containerisierung (Abschnitt 5.4.3) und *Digital Object Identifier* (Abschnitt 5.4.1) stehen Empfehlungen zur Verfügung, um diesem Thema gerecht zu werden.

7.3 ÄNDERUNGEN DER DATENSÄTZE

Das Problem der Auswirkungen von Änderungen an Datenbeständen auf die Reproduzierbarkeit wird in den vorgestellten Lösungsansetzen vermieden, indem dafür gesorgt wird, dass in den bereitgestellten Daten keine Änderungen bezogen auf die Originalarbeit enthalten sind. Dies wird sichergestellt, indem entweder ein Datengenerierungsalgorithmus (Abschnitt 5.2.1) verwendet wird oder indem die Rohdaten, wie sie zum Zeitpunkt der Forschungsarbeit in der Datenbank vorhanden sind, bereitgestellt werden (Abschnitt 5.4.2.1). Ein einfacher Link auf ggf. öffentlich zugängliche Datensätze wird ausdrücklich als nicht ausreichend bezeichnet, eben da sich die Datenquelle ändern oder sogar vollständig verloren gehen kann.

Für den Fall, dass die untersuchte Datenmenge so groß ist, dass es unpraktikabel wäre, den gesamten Datenbestand zu veröffentlichen, es sich aber um echte Datensätze handelt, für die kein Datengenerierungsalgorithmus bereitgestellt werden kann, konnten keine Lösungsansätze gefunden werden.

8 ZUSAMMENFASSUNG UND AUSBLICK

In dieser Seminararbeit wurde der aktuelle Stand und die Herausforderungen des Themas Reproduzierbarkeit für die Datenbanksystem-Forschung analysiert.

In Kapitel 4 wurde festgestellt, dass nur ein kleiner Teil der Forschungsarbeiten im Bereich Informatik und spezieller im Bereich Datenbanksysteme reproduzierbar ist, obwohl dem Thema seit über zehn Jahren Aufmerksamkeit geschenkt wird.

In Kapitel 5 wurden Ansätze angeführt, die für alle in Abschnitt 4.2 angeführten Empfehlungen zur Umsetzung der Reproduzierbarkeitsziele mögliche Lösungen beschreiben.

Zu den in der Einleitung 2.1 gestellten Forschungsfragen wurden die in Kapitel 7 Diskussion genannten Lösungen gefunden. Möglichkeiten für weiterführende Arbeiten ergeben sich aus den offen gebliebenen Fragen bezüglich einer genaueren Untersuchung der Auswirkungen konkreter Benchmarks auf die Reproduzierbarkeit sowie die Frage nach der Praktikabilität der Bereitstellung sehr großer Datenmengen.

Ein weiteres Untersuchungsthema wäre, was gegen die in Abschnitt 4.4 genannten Gründe für Nichtreproduzierbarkeit unternommen werden kann, wenn die in Abschnitt 4.1 genannten Vorteile und Anreize offensichtlich nicht ausreichen, um den Anteil reproduzierbarer Arbeiten zu erhöhen.

Interessant scheinen in diesem Zusammenhang Untersuchungen zur Aufstellung von Empfehlungen für die erwähnte Nutzung von Forschungsarbeiten als Trainingsmaterial für nachfolgende Forschende, denn diese könnte von doppeltem Nutzen sein: Nachwuchsforschende und Studierende könnten ihren Forschungsgegenstand in Form von ausführbaren Papern direkter und nachvollziehbarer kennenlernen als nur durch das Lesen von statischen Papern in PDF-Form. Außerdem könnten sie bereits frühzeitig Methoden guter Forschungspraxis wie die Nutzung von Workflow- und Provenienzsoftware zur Protokollierung von Forschungstätigkeiten sowie die Bereitstellung von Auswertungsskripten in Form von Jupyter-Notebooks kennenlernen. Dieses Wissen kann nicht nur für Forschungspaper, sondern bereits für Studien- und Abschlussarbeiten eingesetzt werden.

Möglicherweise können Bemühungen in diese Richtung dazu beitragen, dem eingangs zitierten Ideal, dass Reproduzierbarkeit fast kein Aufwand sein sollte, näherzukommen.

A

ANHANG

Ergebnisse der Reproduktion des Papers von [DG20].

UDF	Laufzeit				Verbesserung	
	Original		Reproduziert		Original	Repr.
	UDF	compiliert	UDF	compiliert		
comps	357 ms	26 ms	1538 ms	42 ms	7,2 %	2,7 %
eval	216 ms	20 ms	874 ms	29 ms	9,2 %	3,3 %
floyd	>8000 ms	14 ms	>8000 ms	150ms	<1,8 ‰	<0,2 %
fsm	659 ms	102 ms	2084 ms	292 ms	15,4 %	14,0 %
lcs	756 ms	30 ms	4258 ms	30 ms	3,9 %	0,7 %
mandel	280 ms	27 ms	1595 ms	127 ms	9,6 %	7,9 %
march	742 ms	28 ms	2618 ms	22 ms	3,7 %	0,8 %
paths	474 ms	46 ms	1437 ms	135 ms	9,7 %	9,4 %
sizes	144 ms	67 ms	445 ms	202 ms	46,5 %	45,4 %
vm	207 ms	9 ms	838 ms	17 ms	4,3 %	2,0 %

Tabelle A.1: Originale und reproduzierte Laufzeiten der Algorithmen aus Tabelle 2

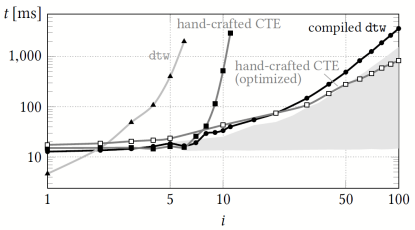


Figure 3: Run time of implementations of `dtw(i,i)`, measured on PostgreSQL 11.3 (logarithmic scales). This work aims for the grey area of low run times.

(a) Originalabbildung 3

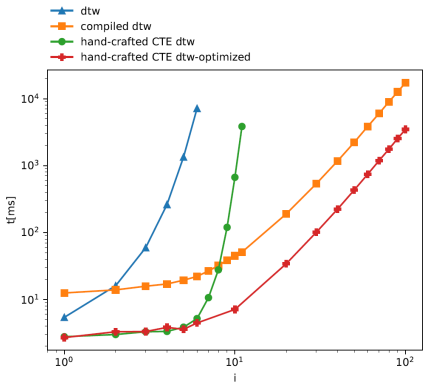


Figure 3: Run time of implementations of dtw(i,i), measured on PostgreSQL 11.3.

(b) Reproduzierte Abbildung 3

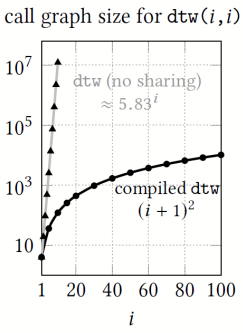


Figure 7: Sharing saves function invocations.

(c) Originalabbildung 7

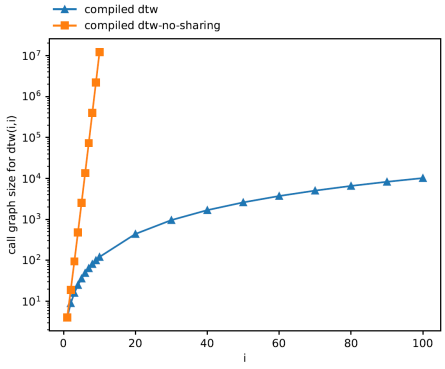
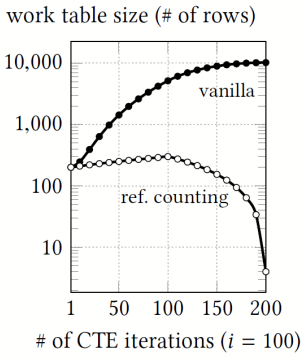


Figure 7: Sharing saves function invocations.

(d) Reproduzierte Abbildung 7



(a) Work table shrinking.

(e) Originalabbildung 20a

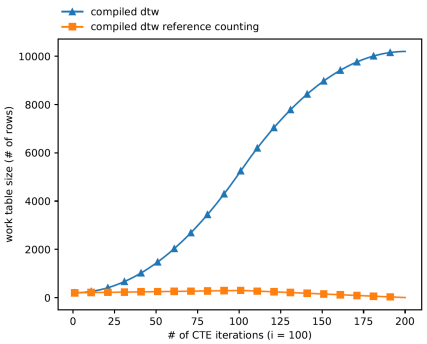
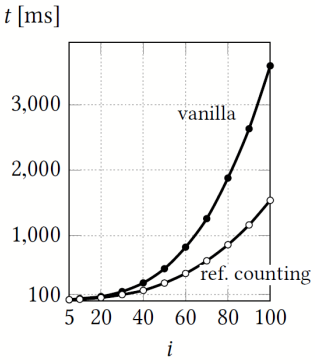


Figure 20 a): Evaluating dtw(i,i): Impact of reference counting on work table size.

(f) Reproduzierte Abbildung 20a



(b) Run time reduction.

(g) Originalabbildung 20b

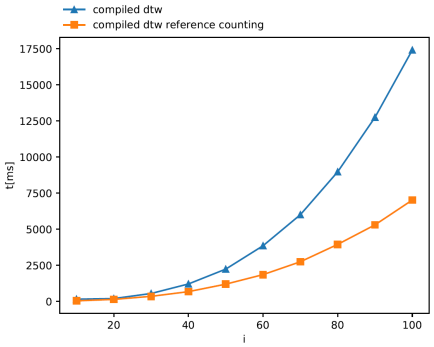


Figure 20 b): Evaluating dtw(i,i): Impact of reference counting on CTE run time.

(h) Reproduzierte Abbildung 20b

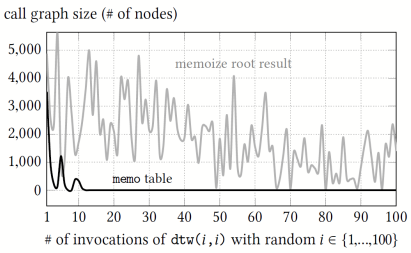


Figure 22: Series of `dtw(i,i)` invocations: Re-using `memo` table entries effectively cuts down call graph size.

(i) Originalabbildung 22

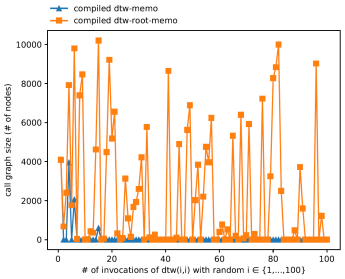


Figure 22: Series of dtw(i,i) invocations: Re-using memotable entries effectively cuts down call graph size.

(j) Reproduzierte Abbildung 22

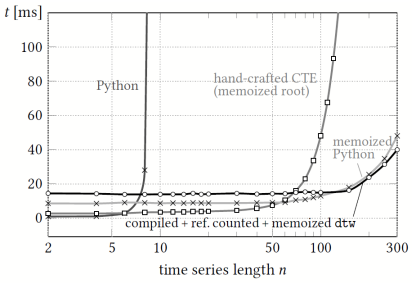


Figure 23: In-database vs. external processing: average evaluation time of `dtw(i,i)` with random $i \in \{1,\ldots,n\}$ (RDBMS: PostgreSQL 11.3, ext. processor: Python 3.7).

(k) Originalabbildung 23

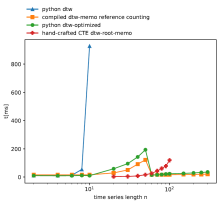


Figure 23: In-database vs. external processing: average evaluation time of dtw(i,i) with random i in {1,...,n} (RDBMS: PostgreSQL 11.3, ext. processor: Python 3.7).

(l) Reproduzierte Abbildung 23

Abbildung A.0: Originalabbildungen und reproduzierte Abbildungen der Arbeit von [DG20]

LITERATUR

[ACM22a] ACM. *Artifact Review and Badging - Current*. Apr. 2022. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current (besucht am 12. 04. 2022).

[ACM22b] ACM. *DB Research Availability & Reproducibility*. 2022. URL: https://reproducibility.sigmod.org/ (besucht am 04. 07. 2022).

[ACM22c] ACM. *Past ACM SIGMOD Reproducibility Efforts*. Mai 2022. URL: https://reproducibility.sigmod.org/history.html (besucht am 25. 05. 2022).

[Bog21] Dmytro Bogatov. "Reproducibility Report for ACM SIGMOD 2020 Paper: "Functional-Style SQL UDFs With a Capital 'F' "". In: *SIGMOD 2021 Reproducibility* (2021).

[Bon+11] Philippe Bonnet u. a. "Repeatability and workability evaluation of SIGMOD 2011". In: *ACM SIGMOD Record* 40.2 (Sep. 2011). Number: 2, S. 45–48. ISSN: 0163-5808. DOI: 10.1145/2034863.2034873. URL: https://doi.org/10.1145/2034863.2034873 (besucht am 11. 04. 2022).

[BMS20] Dimitri Brainingер, Wolfgang Mauerer und Stefanie Scherzinger. "Replicability and Reproducibility of a Schema Evolution Study in Embedded Databases". en. In: *Advances in Conceptual Modeling*. Hrsg. von Georg Grossmann und Sudha Ram. Cham: Springer International Publishing, 2020, S. 210–219. ISBN: 978-3-030-65847-2. DOI: 10.1007/978-3-030-65847-2_19.

[Chi+13] Fernando Chirigati, Matthias Troyer, Dennis Shasha und Juliana Freire. "A Computational Reproducibility Benchmark". In: *IEEE Data Eng. Bull.,* (2013), 36(4):54–59.

[Dat10] Yale Law School Roundtable on Data and Code Sharing. "Reproducible Research". In: *Computing in Science Engineering* 12.5 (Sep. 2010). Number: 5 Conference Name: Computing in Science Engineering, S. 8–13. ISSN: 1558-366X. DOI: 10.1109/MCSE.2010.113.

[DBL22] DBLP-Datenbank. *dblp: computer science bibliography*. 2022. URL: https://dblp.uni-trier.de/ (besucht am 27. 05. 2022).

[Dru12] Dr Chris Drummond. *Reproducible Research: a Dissenting Opinion*. Other. Sep. 2012. URL: https://web-archive.southampton.ac.uk/cogprints.org/8675/ (besucht am 16. 05. 2022).

[DG20] Christian Duta und Torsten Grust. "Functional-Style SQL UDFs With a Capital 'F'". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. Portland OR USA: ACM, Juni 2020, S. 1273–1287. ISBN: 978-1-4503-6735-6. DOI: 10.1145/3318464.3389707. URL: https://dl.acm.org/doi/10.1145/3318464.3389707 (besucht am 22. 06. 2022).

[FBS12a] Juliana Freire, Philippe Bonnet und Dennis Shasha. "Computational reproducibility: state-of-the-art, challenges, and database research opportunities - Tutorial". In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. SIGMOD '12. New York, NY, USA: Association for Computing Machinery, Mai 2012, S. 593–596. ISBN: 978-1-4503-1247-9. DOI: 10.1145/2213836.2213908. URL: https://doi.org/10.1145/2213836.2213908 (besucht am 11. 04. 2022).

[FBS12b] Juliana Freire, Philippe Bonnet und Dennis Shasha. "Computational Reproducibility: State-of-the-art, challenges, database research opportunities". In: SIGMOD '12 (2012).

[GB12] Gabriel F.T. Gomes und Edson Borin. "A Database for Reproducible Computational Research". In: *2012 13th Symposium on Computer Systems*. Okt. 2012, S. 141–147. DOI: 10.1109/WSCAD-SSC.2012.27.

[ISO22] ISO. *ISO 26324:2012*. Mai 2022. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/04/35/43506.html (besucht am 09. 05. 2022).

[LÖ18] Ling Liu und M. Tamer Özsu, Hrsg. *Encyclopedia of Database Systems*. New York, NY: Springer New York, 2018. ISBN: 978-1-4614-8266-6 978-1-4614-8265-9. DOI: 10.1007/978-1-4614-8265-9. URL: http://link.springer.com/10.1007/978-1-4614-8265-9 (besucht am 11. 04. 2022).

[Man+09] Stefan Manegold, Ioana Manolescu, Stefan Manegold und Ioana Manolescu. "Performance evaluation in database research: principles and experience". In: *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. EDBT '09. New York, NY, USA: Association for Computing Machinery, März 2009, S. 1156. ISBN: 978-1-60558-422-5. DOI: 10.1145/1516360.1516503. URL: https://doi.org/10.1145/1516360.1516503 (besucht am 11. 04. 2022).

[MS21] Wolfgang Mauerer und Stefanie Scherzinger. "Nullius in Verba: Reproducibility for Database Systems Research, Revisited". In: *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. ISSN: 2375-026X. Apr. 2021, S. 2377–2380. DOI: 10.1109/ICDE51399.2021.00270.

[Paw+19] Mateusz Pawlik, Thomas Hütter, Daniel Kocher, Willi Mann und Nikolaus Augsten. "A Link is not Enough – Reproducibility of Data". en. In: *Datenbank-Spektrum* 19.2 (Juli 2019). Number: 2, S. 107–115. ISSN: 1610-1995. DOI: 10.1007/s13222-019-00317-8. URL: https://doi.org/10.1007/s13222-019-00317-8 (besucht am 08. 04. 2022).

[Phi22] David Phillips. *electrum/tpch-dbgen*. original-date: 2012-01-18. Mai 2022. URL: https://github.com/electrum/tpch-dbgen (besucht am 30. 05. 2022).

[Ten+11] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff und Mike Frame. "Data sharing by scientists: practices and perceptions". eng. In: *PloS One* 6.6 (2011). Number: 6, e21101. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021101.