„Data Engineering“ in der Hochschullehre

Eine Bestandsaufnahme

Ralf Schenkel¹ · Stefanie Scherzinger² · Marina Tropmann-Frick³

Angenommen: 4. Oktober 2021 / Online publiziert: 29. Oktober 2021
© Der/die Autor(en) 2021

Zusammenfassung

Das Themenheft zu „Data Engineering for Data Science“ gibt uns Anlass, die Rolle dieses Themas in der akademischen Datenbanklehre im Rahmen einer kleinen Umfrage zu erfassen. In diesem Artikel geben wir die Ergebnisse gesammelt wieder. Uns haben 17 Rückmeldungen aus der GI-Fachgruppe Datenbanksysteme erreicht. Im Vergleich zu einer früheren Umfrage zur Lehre im Bereich „Cloud“, 2014 im Datenbankspektrum vorgestellt, zeichnet sich ab, dass Data-Engineering-Inhalte zunehmend auch in grundständigen Lehrveranstaltungen gelehrt werden, sowie außerhalb der Kerninformatik. Data Engineering scheint sich als ein Querschnittsthema zu etablieren, das nicht nur den Masterstudiengängen vorbehalten ist.

Schlüsselwörter Hochschullehre · Data Engineering

1 Einführung

Für den GI-Arbeitskreis „Data Engineering for Data Science“ ist die akademische Ausbildung ein wichtiger Diskussionspunkt. Anlässlich dieses Themenhefts haben wir eine Online-Umfrage durchgeführt, beworben über den E-Mail-Verteiler *fg-db*. Wir haben die Lehrenden im Hochschulbereich um Rückmeldung zu den von ihren Hochschulen angebotenen Data-Engineering-Lehrveranstaltungen gebeten.

Uns erreichten 17 Antworten zu Lehrveranstaltungen an Universitäten sowie Hochschulen für angewandte Wissenschaften. Wir geben diese aggregiert wieder und diskutieren sie im Vergleich zu einer ähnlichen Umfrage aus dem Jahr 2014, zu dem (damals neuen) Thema „Cloud“ [3]. Dabei haben wir bewusst auf eine ähnliche Strukturierung des Online-Fragebogens geachtet, um eine Vergleichbarkeit herstellen zu können.

2 Synopse

Auch wenn unsere Umfrage sich auf eine überschaubare Anzahl Rückmeldungen stützt, so zeigt sich ein deutlicher Konsens bezüglich der Kernaussagen.

Hand-On vs. On-Paper: Die am häufigsten genannten Lehrformate sind Vorlesungen und Praktika am Rechner, letztere dominieren gegenüber dem „Lösen von Aufgaben auf Papier“. Abb. 1 stellt die Rückmeldungen zu Formaten in der aktuellen Umfrage (DE4DS) im Vergleich zu der früheren Umfrage zum Thema Cloud (DMC) dar. Auffallend ist, dass Vorlesungen und Praktika stärker verbreitet sind.

Flächendeckendes Angebot. Data Engineering wird überwiegend als Wahlmodul angeboten, und zwar nicht nur im

---

✉ Ralf Schenkel
schenkel@uni-trier.de

Stefanie Scherzinger
Stefanie.Scherzinger@uni-passau.de

Marina Tropmann-Frick
marina.tropmann-frick@haw-hamburg.de

¹ Professur für Datenbanken und Informationssysteme, Universität Trier, 54296 Trier, Deutschland

² Lehrstuhl für Informatik mit Schwerpunkt skalierbare Datenbanksysteme, Universität Passau, 94032 Passau, Deutschland

³ Professur für Datenanalyse, HAW Hamburg, 20099 Hamburg, Deutschland

Abb. 1 Formate bei Data-Engineering-Lehrveranstaltungen (DE4DS), im Vergleich mit einer früheren Umfrage zum Thema „Cloud“ (DMC). V: Vorlesung, P: Praktikum, S: Seminar

Abb. 2 Verankerung in Informatik-Studiengängen: Wahlpflichtfach oder Pflichtfach im Master (WMSc vs. PMSc), bzw. im Bachelor (WBSc vs. PBSc)

Master Informatik, sondern auch vermehrt im Bachelor Informatik, wie in Abb. 2 visualisiert.

Laut unseren Umfrageergebnissen sind die meisten Veranstaltungen eher klein konzipiert, mit bis zu 50 Studierenden. Das lässt auf einen guten Betreuungsschlüssel schließen. Damit bleibt der openHPI-MOOC-Kurs von Felix Naumann, mit einer Reichweite von zehntausend Personen [2], außer Konkurrenz.

Inhalte. Unter den erwarteten fachlichen Voraussetzungen wurden am häufigsten Datenbank-Kenntnisse sowie die Beherrschung der Programmiersprachen Python und Java genannt. Hier müssen wir von einem *Bias* bei dieser Umfrage ausgehen, wodurch die Datenbank-Kenntnisse wahrscheinlich betont wurden.

Dafür wurden Statistik-Kenntnisse selten genannt. Interdisziplinäre Themen, wie Datenethik, Privacy oder juristische Aspekte, wurden gar nicht genannt. Wir greifen diese Beobachtungen in der Diskussion auf.

Bei der Frage nach den vermittelten Methoden wurden an erster Stelle Big-Data-Technologien und Datenintegra-

tion genannt, an zweiter Stelle Data Mining, und an dritter Stelle Methoden des maschinellen Lernens. Auch Herausforderungen wie die Skalierbarkeit, oder die Verarbeitung von Graphen, fanden Erwähnung.

Sprache je nach Studienphase. Als Lehrsprachen wurden Deutsch und Englisch zu gleichen Teilen genannt. Wir vermuten, dass dies der typischen Aufteilung „Deutsch im Bachelor“ und „Englisch im Master“ folgt. Diese Zuordnung könnte man bei einer detaillierteren Untersuchung erfragen.

Taktung. In den Antworten zu unserer Umfrage wurde von einer Veranstaltung berichtet, die bereits seit 15 Jahren angeboten wird. Die meisten vorgestellten Veranstaltungen sind jedoch seit ca. 6 Jahren im Angebot, im jährlichen Turnus.

Export-Schlager. Bei der Frage, für welche Studiengänge die Veranstaltungen angeboten werden, erweist sich Data Engineering als Exportschlager. So zeigte die Befragung den Export in eigene Data-Science-Studiengänge, gefolgt von Wirtschaftsinformatik. Data Engineering wird breit exportiert, auch in die Mathematik und die Naturwissenschaften hinein.

Lehrmaterial. Über Lehrmaterial wurde berichtet, dass überwiegend eigene Folien zur Verfügung gestellt werden, aber auch, dass „Videos im Kommen“ seien (was natürlich gegenwärtig auch durch die pandemische Lage verursacht wird).

Beim Einsatz von Literatur finden sich Gemeinsamkeiten mit der früheren Umfrage: Sachbücher dominieren über Lehrbücher. Ein möglicher Grund ist, dass sie viel schneller auf dem Markt sind.

IT-Infrastruktur. Bei der den Studierenden angebotenen IT-Infrastruktur wurde die „Lehrstuhl-Cloud“ genannt, gefolgt von virtuellen Maschinen, die auf den privaten Rechnern der Studierenden laufen. Uns hat fünf Mal die Antwort erreicht, dass den Studierenden keine Infrastruktur zur Verfügung gestellt wird, d.h., die Studierenden richten sich ihre IT-Umgebung autark ein.

Bei den verwendeten Softwareprodukten herrscht Vielfalt, insbesondere findet sich Vieles aus dem Apache-Portfolio. Python wurde mehrfach genannt, Jupyter Notebooks nur einmal, R oder Julia wurden nicht erwähnt.

In den Freitextkommentaren wurde angegeben, dass Hadoop als System immer weniger relevant ist, während Cloud-Technologien weiterhin an Stellenwert gewinnen. Das spiegelt auch unsere eigene Beobachtung wieder.

3 Diskussion

Unsere Umfrage zeigt, dass die Data-Engineering-Lehre vergleichsweise stark in die Breite geht, in die Bachelorstudiengänge hinein, und auch in Studiengängen jenseits der Kern-Informatik. Da Data Science immer eine Anwendungsdomäne benötigt (bzw. einen „Bindestrich“), scheint dies nur natürlich.

Betrachten wir die Rückmeldungen im Kontext der Empfehlungen der GI für Data-Science-Studiengänge [1] (im Folgeheft werden diese in einem eigenen Artikel vorgestellt), so fällt auf, dass keine Lehrinhalte zu Statistik genannt wurden. Hier können wir vermuten, dass diese Inhalte nicht von den Kollegen und Kolleginnen mit Datenbank-Expertise gelehrt werden. Andere zentrale Inhalte, wie etwa Ethik im Umgang mit Daten, vermuten wir ebenso in anderen Lehrveranstaltungen.

Zwar scheint „Cloud-Technologie“ als Lehrinhalt weiterhin präsent, doch werden kommerzielle Cloud-Dienste für die Lehre selten genutzt — in unserer Umfrage zur verwendeten IT-Infrastruktur wurde diese Option nur einmal genannt. In der früheren Umfrage zum Thema „Cloud“ war diese Option noch prominenter vertreten.

Wir vermuten als einen Grund die anhaltende Schwierigkeit, kostenpflichtige Cloud-Dienste in die Lehre zu integrieren. Der Erfahrung des Autorenteams nach hat sich in Sachen Rechtssicherheit und praktikable Rechnungslegung nicht viel getan: Die Probleme, die 2014 beschrieben wurden [3], sind an vielen Hochschulen weiterhin ungelöst. Hier sehen wir dringenden Handlungsbedarf, um zu ermöglichen, dass Cloud-Dienste in der Hochschullehre niederschwellig integrierbar sind.

Danksagung Wir danken allen, die an unserer Umfrage teilgenommen haben. Wir danken auch Meike Klettke und Uta Störl für ihr Feedback zu diesem Artikel.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

1. Abedjan Z, Bendig T, Brefeld U, Bürkle J, Desel J, Edlich S, Eppler T, Goedicke M, Hachmeister N, Heidrich J, Höppner S, Kast SM, Krupka D, Lang K, Liggesmeyer P, Meisner J, Scholtes I, Tropmann-Frick M (2021) Empfehlungen für Masterstudiengänge „Data Science“ – auf Basis eines Bachelors in (Wirtschafts-)Informatik oder Mathematik. <https://dl.gi.de/handle/20.500.12116/36788>. Zugriffen: 27. Sept. 2021
2. Alder N, Bleifuß T, Bornemann L, Naumann F, Repke T (2021) Ein Data Engineering Kurs für 10.000 Teilnehmer. Datenbank Spektrum 21(1):5–9. <https://doi.org/10.1007/s13222-020-00354-8>
3. Scherzinger S, Thor A (2014) Cloud-Technologien in der Hochschullehre – Pflicht oder Kür? – Eine Standortbestimmung innerhalb der GI-Fachgruppe Datenbanksysteme. Datenbank Spektrum 14(2):131–134. <https://doi.org/10.1007/s13222-014-0161-2>