

Founding Editors

Gerhard Goos

Karlsruhe Institute of Technology, Karlsruhe, Germany

Juris Hartmanis

Cornell University, Ithaca, NY, USA

Editorial Board Members

Elisa Bertino

Purdue University, West Lafayette, IN, USA

Wen Gao

Peking University, Beijing, China

Bernhard Steffen 

TU Dortmund University, Dortmund, Germany

Gerhard Woeginger 

RWTH Aachen, Aachen, Germany

Moti Yung

Columbia University, New York, NY, USA

More information about this subseries at <http://www.springer.com/series/7410>

Vijay Gadepally · Timothy Mattson ·
Michael Stonebraker · Tim Kraska ·
Fusheng Wang · Gang Luo ·
Jun Kong · Alevtina Dubovitskaya (Eds.)


Heterogeneous Data Management, Polystores, and Analytics for Healthcare

VLDB Workshops, Poly 2020 and DMAH 2020
Virtual Event, August 31 and September 4, 2020
Revised Selected Papers

Editors

Vijay Gadepally
Massachusetts Institute of Technology
Lexington, MA, USA

Michael Stonebraker
Massachusetts Institute of Technology
Cambridge, MA, USA

Fusheng Wang 
Stony Brook University
Stony Brook, NY, USA

Jun Kong
Georgia State University
Atlanta, GA, USA

Timothy Mattson
Intel Corporation
Portland, OR, USA

Tim Kraska
Massachusetts Institute of Technology
Cambridge, MA, USA

Gang Luo
University of Washington
Seattle, WA, USA

Alevtina Dubovitskaya
Lucerne University of Applied Sciences
Rotkreuz, Switzerland

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-030-71054-5

ISBN 978-3-030-71055-2 (eBook)

<https://doi.org/10.1007/978-3-030-71055-2>

LNCS Sublibrary: SL4 – Security and Cryptology

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In this volume we present the accepted contributions for the VLDB conference workshops entitled: Polystore systems for heterogeneous data in multiple databases with privacy and security assurances (Poly'20) and the Sixth International Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH 2020) held virtually with the 46th [International Conference on Very Large Data Bases](#) on August 31 and September 4, 2020.

Poly'20 Overview:

Enterprises are routinely divided into independent business units to support agile operation. However, this leads to “siloeed” information systems. Such silos generate a host of problems, such as:

DISCOVERY of relevant data to a problem at hand. For example: Merck has 4000 (+/-) Oracle databases, a data lake, large numbers of files and an interest in public data from the web. Finding relevant data in this sea of information is a challenge.

INTEGRATING the discovered data. Independently constructed schemas are never compatible.

CLEANING the resulting data. A good figure of merit is that 10% of all data is missing or wrong.

ENSURING EFFICIENT ACCESS to resulting data. At scale operations must be performed “in situ”, and a good polystore system is a requirement.

It is often said that data scientists spend 80% (or more) of their time on these tasks, and it is crucial to have better solutions.

In addition, the EU has recently enacted GDPR that will force enterprises to assuredly delete personal data on request. This “right to be forgotten” is one of several requirements of GDPR, and it is likely that GDPR-like requirements will spread to other locations, for example California. In addition, privacy and security issues are increasingly an issue for large internet platforms. In enterprises, these issues will be front and center in the distributed information systems in place today.

Lastly, enterprise access to data in practice will require queries constructed from a variety of programming models. A “one size fits all” [1] mentality just won't work in these cases.

DMAH'20 Overview:

The goal of the workshop is to bring together researchers from the cross-cutting domains of research including information management and biomedical informatics. The workshop aims to foster exchange of information and discussions on innovative data management and analytics technologies. We encourage topics that highlight end-to-end applications, systems, and methods addressing problems in healthcare, public health, and everyday wellness; integration with clinical, physiological, imaging, behavioral, environmental, and “omics” data, as well as data from social media and the Web. Our hope for this workshop is to provide a unique opportunity for mutually beneficial and informative interaction between information management and biomedical researchers from interdisciplinary fields.

Organization

POLY'20

Workshop Chairs

Vijay Gadepally	Massachusetts Institute of Technology, USA
Tim Kraska	Massachusetts Institute of Technology, USA
Timothy Mattson	Intel Corporation, USA
Michael Stonebraker	Massachusetts Institute of Technology, USA

Program Committee Members

Danny Weitzner	MIT Internet Policy Research Initiative, USA
Michael Gubanov	Florida State University, USA
Edmon Begoli	Oak Ridge National Laboratory, USA
Dimitris Kolovos	University of York, UK
Amarnath Gupta	University of California, San Diego, USA
Ratnesh Sahay	AstraZeneca, UK
Rada Chirkova	North Carolina State University, USA
Sam Madden	Massachusetts Institute of Technology, USA
Pedro Pedreira	Facebook Inc., USA
Makoto Onizuka	University of Osaka, Japan

DMAH 2020

Workshop Chairs

Fusheng Wang	Stony Brook University, USA
Gang Luo	University of Washington, USA
Jun Kong	Georgia State University, USA
Alevtina Dubovitskaya	Lucerne University of Applied Sciences and Arts and Swisscom, Switzerland

Program Committee Members

Edmon Begoli	Oak Ridge National Laboratory, USA
Yang Cao	Kyoto University, Japan
Blair Christian	Oak Ridge National Laboratory, USA
Dejing Dou	University of Oregon, USA
Alevtina Dubovitskaya	Lucerne University of Applied Sciences and Arts and Swisscom, Switzerland

Peter Elkin	University at Buffalo, USA
Zhe He	Florida State University, USA
Vagelis Hristidis	University of California, Riverside, USA
Athirai Irissappane	University of Washington, USA
Guoqian Jiang	Mayo Clinic College of Medicine and Science, USA
Jun Kong	Georgia State University, USA
Tahsin Kurc	Stony Brook University, USA
Yanhui Liang	Google Inc., USA
Gang Luo	University of Washington, USA
Fusheng Wang	Stony Brook University, USA
Ye Ye	University of Pittsburgh, USA
Rui Zhang	University of Minnesota, USA

Using Demographic Pattern Analysis to Predict COVID-19 Fatalities on the US County Level (Abstract of DMAH 2020 Invited Talk)

Klaus Muller

Stony Brook University, Stony Brook, New York, USA
Akai Kaeru LLC, New York, New York, USA
mueller@cs.stonybrook.edu

Abstract. Unlike pandemics in the past, COVID-19 has hit us in the midst of the information age. We have built vast capabilities to collect and store data of any kind which can be analyzed in myriad ways to help us mitigate the impact of this catastrophic disease. Specifically for COVID-19, data analysis can help local governments to plan the allocation of testing kits, testing stations, and primary care units, and it can help them in setting guidelines for residents, such as the need for social distancing, the use of face masks, and when to open local businesses that enable human contact. Further, it can also lead to a better understanding of pandemics in general and so inform policy makers on the regional and national level. All of this can save both cost and lives. In this talk I will present the results of an ongoing study we conducted using a prominent regularly updated dataset. We used a pattern mining engine we developed to find specific characteristics of US counties that appear to expose them to higher COVID-19 mortality. Furthermore, we also show that these characteristics can be used to predict future COVID-19 mortality.

Bio. Dr. Klaus Mueller received a PhD in computer science from The Ohio State University. He is currently a professor in the Computer Science Department at Stony Brook University and he is also a senior scientist at the Computational Science Initiative at Brookhaven National Lab. He is also a co-founder of Akai Kaeru, the start-up where most of this research took place. His current research interests are explainable AI, visual analytics, and data science. He won the US National Science Foundation Early Career Award, the SUNY Chancellor Award for Excellence in Scholarship and Creative Activity, and the Meritorious Service Certificate and the Golden Core Award of the IEEE Computer Society. Klaus was inducted into the National Academy of Inventors. To date, he has authored more than 200 peer-reviewed journal and conference papers, which have been cited more than 10,000 times. He is a frequent speaker at international conferences, has organized or participated in 18 tutorials on various topics, chaired the IEEE Visualization Conference, and was the elected chair of the IEEE Technical Committee on Visualization and Computer Graphics (VGTC). Klaus currently serves as the Editor-in-Chief of IEEE Transactions on Visualization and Computer Graphics. He is a senior member of the IEEE.

Contents

Poly 2020: Privacy, Security and/or Policy Issues for Heterogenous Data

A Polystore Based Database Operating System (DBOS).	3
<i>Michael Cafarella, David DeWitt, Vijay Gadepally, Jeremy Kepner, Christos Kozyrakis, Tim Kraska, Michael Stonebraker, and Matei Zaharia</i>	
Polypheny-DB: Towards Bridging the Gap Between Polystores and HTAP Systems	25
<i>Marco Vogt, Nils Hansen, Jan Schönholz, David Lengweiler, Isabel Geissmann, Sebastian Philipp, Alexander Stiemer, and Heiko Schuldt</i>	
Persona Model Transfer for User Activity Prediction Across Heterogeneous Domains.	37
<i>Takahiro Hara</i>	
PolyMigrate: Dynamic Schema Evolution and Data Migration in a Distributed Polystore.	42
<i>Alexander Stiemer, Marco Vogt, Heiko Schuldt, and Uta Störl</i>	
An Architecture for the Development of Distributed Analytics Based on Polystore Events.	54
<i>Athanasios Zolotas, Konstantinos Barmpis, Fady Medhat, Patrick Neubauer, Dimitris Kolovos, and Richard F. Paige</i>	
Towards Data Discovery by Example	66
<i>El Kindi Rezig, Allan Vanterpool, Vijay Gadepally, Benjamin Price, Michael Cafarella, and Michael Stonebraker</i>	
The Transformers for Polystores - The Next Frontier for Polystore Research.	72
<i>Edmon Begoli, Sudarshan Srinivasan, and Maria Mahbub</i>	
DMAH 2020: COVID-19 Data Analytics and Visualization	
Open-World COVID-19 Data Visualization [Extended Abstract]	81
<i>Hyunseung Hwang and Steven Euijong Whang</i>	

DMAH 2020: Deep Learning based Biomedical Data Analytics

Privacy-Preserving Knowledge Transfer with Bootstrap Aggregation
of Teacher Ensembles 87
*Hong-Jun Yoon, Hilda B. Klasky, Eric B. Durbin, Xiao-Cheng Wu,
Antoinette Stroup, Jennifer Doherty, Linda Coyle, Lynne Penberthy,
Christopher Stanley, J. Blair Christian, and Georgia D. Tourassi*

An Intelligent and Efficient Rehabilitation Status Evaluation Method:
A Case Study on Stroke Patients. 100
Yao Tong, Hang Yan, Xin Li, Gang Chen, and Zhenxiang Zhang

Multiple Interpretations Improve Deep Learning Transparency for Prostate
Lesion Detection. 120
Mehmet A. Gulum, Christopher M. Trombley, and Mehmed Kantardzic

DMAH 2020: NLP Based Learning from Unstructured Data

Tracing State-Level Obesity Prevalence from Sentence Embeddings
of Tweets: A Feasibility Study 141
Xiaoyi Zhang, Rodoniki Athanasiadou, and Narges Razavian

Enhancing Medical Word Sense Inventories Using Word Sense Induction:
A Preliminary Study 151
Qifei Dong and Yue Wang

DMAH 2020: Biomedical Data Modelling and Prediction

Teaching Analytics Medical-Data Common Sense. 171
Tomer Sagi, Nitzan Shmueli , Bruce Friedman, and Ruth Bergman

CDRGen: A Clinical Data Registry Generator
(Formal and/or Technical Paper) 188
*Pedro Alves, Manuel J. Fonseca, João D. Pereira,
and Helena Galhardas*

Prediction of lncRNA-Disease Associations from Tripartite Graphs 205
Mariella Bonomo, Armando La Placa, and Simona E. Rombo

DMAH 2020: Invited Paper

Parameter Sensitivity Analysis for the Progressive Sampling-Based
Bayesian Optimization Method for Automated Machine Learning
Model Selection 213
Weipeng Zhou and Gang Luo

Short Paper

Extended Abstract: Programming Heterogeneous Data Applications with Knowledge Graphs.	231
<i>Michael Cafarella</i>	
Author Index	233