TABLE OF CONTENTS

Front Matter

Copyright Notice ............................................................................................................................................................... i
Table of Contents .............................................................................................................................................................. ii
PVLDB Organization and Review Board – Vol. 16 .................................................................................................... vi

Research Papers

A Two-Level Signature Scheme for Stable Set Similarity Joins................................................................................... 2686
Daniel Schmitt, Daniel Kocher, Nikolaus Augsten, Willi Mann, Alexander Miller

Scalable Reasoning on Document Stores via Instance-Aware Query Rewriting..................................................... 2699
Olivier Rodriguez, Federico Ulliana, Marie-Laure Mugnier

EQUI-VOCAL: Synthesizing Queries for Compositional Video Events from Limited User Interactions........ 2714
Enhao Zhang, Maureen Daum, Dong He, Brandon Haynes, Ranjay Krishna, Magdalena Balazinska

Lotan: Bridging the Gap between GNNs and Scalable Graph Analytics Engines ................................................... 2728
Yuhao Zhang, Arun Kumar

Epoxy: ACID Transactions Across Diverse Data Stores ........................................................................................... 2742
Peter Kraft, Qian Li, Xinqing Zhou, Peter Bailis, Michael Stonebraker, Xiayao Yu, Matei Zaharia

Analyzing Vectorized Hash Tables Across CPU Architectures ................................................................................... 2755
Maximilian Böther, Lawrence Benson, Ana Klimovic, Tilmann Rabl

Exploiting Cloud Object Storage for High-Performance Analytics ............................................................................ 2769
Dominik Durner, Viktor Leis, Thomas Neumann

A Randomized Blocking Structure for Streaming Record Linkage ............................................................................. 2783
Dimitrios Karapiperis, Christos Tjortjis, Vassilios S. Verykios

REmatch: a novel regex engine for finding all matches ................................................................................................ 2792
Cristian Riveros, Nicolás Van Sint Jan, Domagoj Vrgoč

ADOPT: Adaptively Optimizing Attribute Orders for Worst-Case Optimal Join Algorithms via Reinforcement Learning.................................................................................................................................................................................... 2805
Junxiong Wang, Immanuel Trummer, Ahmet Kara, Dan Olteanu

Triangular Stability Maximization by Influence Spread over Social Networks ....................................................... 2818
Zheng Hu, Weiguo Zheng, Xiang Lian

CORE-Sketch: On Exact Computation of Median Absolute Deviation with Limited Space .................................. 2832
Haoquan Guan, Ziling Chen, Shaoku Song

Fast Search-By-Classification for Large-Scale Databases Using Index-Aware Decision Trees and Random Forests .................................................................................................................................................................................... 2845
Christian Lülf, Denis Mayr Lima Martins, Marcos Antonio Vaz Salles, Yongluan Zhou, Fabian Gieseke

Semi-Oblivious Chase Termination for Linear Existential Rules: An Experimental Study .................................. 2858
Marco Calautti, Mostafa Milani, Andreas Pieris

Analyzing the Impact of Cardinality Estimation on Execution Plans in Microsoft SQL Server .......................... 2871
Kukjin Lee, Anshuman Dutt, Vivek Narasayya, Surajit Chaudhuri
WALTZ: Leveraging Zone Append to Tighten the Tail Latency of LSM Tree on ZNS SSD ................................. 2884
Jongsung Lee, Donguk Kim, Jae W. Lee

Accelerating Aggregation Queries on Unstructured Streams of Data ......................................................... 2897
Matthew Russo, Tatsunori Hashimoto, Daniel Kang, Yi Sun, Matei Zaharia

QueryBooster: Improving SQL Performance Using Middleware Services for Human-Centered Query Rewriting ................................................................. 2911
Qiushi Bai, Sadeem Alsudais, Chen Li

Consistent Range Approximation for Fair Predictive Modeling ....................................................................... 2925
Jiongli Zhu, Sainyam Galhotra, Nazanin Sabri, Babak Salimi

SUREL+: Moving from Walks to Sets for Scalable Subgraph-based Graph Representation Learning ............... 2939
Haoteng Yin, Muhan Zhang, Jianguo Wang, Pan Li

Estimating Single-Node PageRank in $\tilde{O}(\min\{d, \sqrt{m}\})$ Time ..................................................... 2949
Hanzhi Wang, Zhewei Wei

Simple Adaptive Query Processing vs. Learned Query Optimizers: Observations and Analysis ...................... 2962
Yunjia Zhang, Yannis Chronis, Jignesh M. Patel, Theodoros Rekatsinas

BP-tree: Overcoming the Point-Range Operation Tradeoff for In-Memory B-trees ........................................ 2976
Helen Xu, Amanda Li, Brian Wheatman, Manoj Marneni, Prashant Pandey

HENCE-X: Toward Heterogeneity-agnostic Multi-level Explainability for Deep Graph Networks ................. 2990
Ge Lv, Chen Jason Zhang, Lei Chen

Automatic Road Extraction with Multi-Source Data Revisited: Completeness, Smoothness and Discrimination .......................................................... 3004
Haitao Yuan, Sai Wang, Zhifeng Bao, Shangguang Wang

Asymptotically Better Query Optimization Using Indexed Algebra ............................................................. 3018
Philipp Fent, Guido Moerkotte, Thomas Neumann

Normalizing Property Graphs ......................................................................................................................... 3031
Philipp Skavantzos, Sebastian Link

A Deep Dive into Common Open Formats for Analytical DBMSs ................................................................. 3044
Chunwei Liu, Anna Pavlenko, Matteo Interlandi, Brandon Haynes

Saibot: A Differentially Private Data Search Platform .................................................................................. 3057
Zezhou Huang, Jiaxiang Liu, Daniel Gbenga Alabi, Raul Castro Fernandez, Eugene Wu

JoinBoost: Grow Trees Over Normalized Data Using Only SQL ................................................................. 3071
Zezhou Huang, Rathijit Sen, Jiaxiang Liu, Eugene Wu

R$^3$: Record-Replay-Retroaction for Database-Backed Applications ............................................................. 3085
Qian Li, Peter Kraft, Michael Cafarella, Çağatay Demiralp, Goetz Graefe, Christos Kozyrakis, Michael Stonebraker, Lalith Suresh, Xiangyao Yu, Matei Zaharia

Self-Training for Label-Efficient Information Extraction from Semi-Structured Web-Pages ......................... 3098
Ritesh Sarkhel, Binxuan Huang, Colin Lockard, Prashant Shiralkar

Efficient Non-Learning Similar Subtrajectory Search .................................................................................... 3111
Jiabao Jin, Peng Cheng, Lei Chen, Xuemin Lin, Wenjie Zhang
Frequency-revealing attacks against Frequency-hiding Order-preserving Encryption ........................................ 3124
Xinle Cao, Jian Liu, Yongsheng Shen, Xiaohua Ye, Kui Ren

Efficient Fault Tolerance for Recommendation Model Training via Erasure Coding ........................................ 3137
Tianyu Zhang, Kaige Liu, Jack Kosaian, Juncheng Yang, Rashmi Vinayak

SlabCity: Whole-Query Optimization using Program Synthesis ................................................................. 3151
Rui Dong, Jie Liu, Yuxuan Zhu, Cong Yan, Barzan Mozafari, Xinyu Wang

Scaling Up Structural Clustering to Large Probabilistic Graphs Using Lyapunov Central Limit Theorem ..... 3165
Joseph Howie, Venkatesh Srinivasan, Alex Thomo

Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy ........................................ 3178
Lucas Rosenblatt, Bernease Herman, Anastasia Holovenko, Wonkwon Lee, Joshua Loftus, Elizabeth Mckinnie,
Taras Rumezhak, Andrii Studnik, Bill Howe, Julia Stoyanovich

POEM: Pattern-Oriented Explanations of Convolutional Neural Networks .................................................. 3192
Vargha Dadvar, Lukasz Golab, Divesh Srivastava

gCore: Exploring Cross-layer Cohesiveness in Multi-layer Graphs .............................................................. 3201
Dandan Liu, Zhaonian Zou

Sieve: A Learned Data-Skipping Index for Data Analytics ................................................................. 3214
Yulai Tong, Jiazhen Liu, Hua Wang, Ke Zhou, Rongfeng He, Qin Zhang, Cheng Wang

Out-of-Order Sliding-Window Aggregation with Efficient Bulk Evictions and Insertions .............................. 3227
Kanat Tangwongsan, Martin Hirzel, Scott Schneider

k-Best Egalitarian Stable Marriages for Task Assignment ........................................................................ 3240
Siyuan Wu, Leong Hou U, Panagiotis Karras

Federated Calibration and Evaluation of Binary Classifiers ........................................................................... 3253
Graham Cormode, Igor L. Markov

FlashAlloc: Dedicating Flash Blocks By Objects ............................................................................................. 3266
Jonghyeok Park, Soyee Choi, Gihwan Oh, Soojun Im, Moon-Wook Oh, Sang-Won Lee

Through the Fairness Lens: Experimental Analysis and Evaluation of Entity Matching ................................. 3279
Nima Shahbazi, Nikola Danevski, Fatemeh Nargesian, Abolfazl Asudeh, Divesh Srivastava

Check Out the Big Brain on BRAD: Simplifying Cloud Data Processing with Learned Automated Data Meshes ................................................................................................................................................................................. 3293
Tim Kraska, Tianyu Li, Samuel Madden, Markus Markakis, Amadou Ngom, Ziniu Wu, Geoffrey X. Yu

How Large Language Models Will Disrupt Data Management ..................................................................... 3302
Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, Chenhao Tan

FASTgres: Making Learned Query Optimizer Hinting Effective ...................................................................... 3310
Lucas Woltmann, Jerome Thiessat, Claudio Hartmann, Dirk Habich, Wolfgang Lehner

Write-Aware Timestamp Tracking: Effective and Efficient Page Replacement for Modern Hardware .............. 3323
Demian Vöhringer, Viktor Leis
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tigger: A Database Proxy That Bounces With User-Bypass</td>
<td>3335</td>
</tr>
<tr>
<td>Matthew Butrovich, Karthik Ramanathan, John Rollinson, Wan Shen Lim, William Zhang, Justine Sherry, Andrew Pavlo</td>
<td></td>
</tr>
<tr>
<td>Equitable Data Valuation Meets the Right to Be Forgotten in Model Markets</td>
<td>3349</td>
</tr>
<tr>
<td>Haocheng Xia, Jinfel Liu, Jian Lou, Zhan Qin, Kui Ren, Yang Cao, Li Xiong</td>
<td></td>
</tr>
<tr>
<td>TSM-Bench: Benchmarking Time Series Database Systems for Monitoring Applications</td>
<td>3363</td>
</tr>
<tr>
<td>Abdelouahab Khelifati, Mourad Khayati, Anton Dignös, Djellel Difallah, Philippe Cudré-Mauroux</td>
<td></td>
</tr>
<tr>
<td>Cross Modal Data Discovery over Structured and Unstructured Data Lakes</td>
<td>3377</td>
</tr>
<tr>
<td>Mohamed Y. Eltabakh, Mayuresh Kunjir, Ahmed K. Elmagarmid, Mohammad Shahmeer Ahmad</td>
<td></td>
</tr>
<tr>
<td>Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples</td>
<td>3391</td>
</tr>
<tr>
<td>Peng Li, Yeye He, Cong Yan, Yue Wang, Surajit Chaudhuri</td>
<td></td>
</tr>
<tr>
<td>Effective Entity Augmentation By Querying External Data Sources</td>
<td>3404</td>
</tr>
<tr>
<td>Christopher Buss, Jasmin Mousavi, Mikhail Tokarev, Arash Termehchy, David Maier, Stefan Lee</td>
<td></td>
</tr>
<tr>
<td>Choose Wisely: An Extensive Evaluation of Model Selection for Anomaly Detection in Time Series</td>
<td>3418</td>
</tr>
<tr>
<td>Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, Themis Palpanas</td>
<td></td>
</tr>
<tr>
<td>Similarity search in the blink of an eye with compressed indices</td>
<td>3433</td>
</tr>
<tr>
<td>Cecilia Aguerrebere, Ishwar Singh Bhati, Mark Hildebrand, Mariano Tepper, Theodore Willke</td>
<td></td>
</tr>
<tr>
<td>On Data-Aware Global Explainability of Graph Neural Networks</td>
<td>3447</td>
</tr>
<tr>
<td>Ge Lv, Lei Chen</td>
<td></td>
</tr>
<tr>
<td>Declarative Sub-Operators for Universal Data Processing</td>
<td>3461</td>
</tr>
<tr>
<td>Michael Jungmair, Jana Giceva</td>
<td></td>
</tr>
</tbody>
</table>
PVLDB ORGANIZATION AND REVIEW BOARD - Vol. 16

Editors in Chief of PVLDB
Georgia Koutrika (Athena Research Center)
Jun Yang (Duke University)

Associate Editors of PVLDB
Alkis Simitsis (Athena Research Center)
Amol Deshpande (University of Maryland at College Park)
Angela Bonifati (Lyon 1 University)
Ashwin Machanavajjhala (Duke University/Tumult Labs)
Badrish Chandramouli (Microsoft Research)
Boris Glavic (Illinois Institute of Technology)
Ce Zhang (ETH Zurich)
Cyrus Shahabi (University of Southern California)
Dan Olteanu (University of Zurich)
Eric Lo (The Chinese University of Hong Kong)
Evaggelia Pitoura (University of Ioannina)
Evimaria Terzi (Boston University)
Gustavo Alonso (ETH Zurich)
Helen (Zi) Huang (University of Queensland)
Hong Cheng (The Chinese University of Hong Kong)
Kenneth Ross (Columbia University)
Khuzaima Daudjee (University of Waterloo)
Kyuseok Shim (Seoul National University)
Letizia Tanca (Politecnico di Milano)
Lucian Popa (IBM Research - Almaden)
Magdalena Balazinska (University of Washington)
Meihui Zhang (Beijing Institute of Technology)
Neo Ziss Polyzotsis (Databricks)
Nesime Tatbul (Intel Labs and MIT)
Nikos Mamoulis (University of Ioannina)
Rachel Pottinger (University of British Columbia)
Wenjie Zhang (University of New South Wales)
Wolfgang Gatterbauer (Northeastern University)
Wook-Shin Han (Pohang University of Science and Technology)
Xiaokui Xiao (National University of Singapore)
Yannis Velegrakis (University of Trento and Utrecht University)
Yanyan Shen (Shanghai Jiao Tong University)
Yi Chen (New Jersey Institute of Technology)
Yongxin Tong (Beihang University)
Zhifeng Bao (RMIT University)

Publication Editors
Manos Athanassoulis (Boston University)
Kostas Stefanidis (Tampere University)
Ju Fan (Renmin University of China)

PVLDB Managing Editor
Wolfgang Lehner (Dresden University of Technology)

PVLDB Advisory Board
Vanessa Braganholo (Universidade Federal Fluminense)
Sourav S Bhowmick (Nanyang Technological University)
Chris Jermaine (Rice University)
Peter Triantafillou (University of Warwick)
Xin Luna Dong (Facebook)
Fatma Ozcan (Google)
Lei Chen (Hong Kong University of Science and Technology)
Juliana Freire (New York University)
Graham Cormode (University of Warwick)
Divesh Srivastava (AT&T Labs-Research)
Wolfgang Lehner (Dresden University of Technology)
Felix Naumann (HPI)
Xuemin Lin (University of New South Wales)
Georgia Koutrika (Athena Research Center)
Jun Yang (Duke University)
Review Board

Abolfazl Asudeh (University of Illinois at Chicago)
Alexander J Ratner (University of Washington)
Alexandra Meliou (University of Massachusetts Amherst)
Amelie Marian (Rutgers University)
Amir Gilad (Duke University)
Amir Shaikhha (University of Edinburgh)
Anastasiou Kementsietsidis (Google Research)
Andrew Crotty (Carnegie Mellon University)
Anna Fariha (Microsoft)
Anton Dignós (Free University of Bozen-Bolzano)
Antonios Deligiannakis (Technical University of Crete)
Arijit Khan (Nanyang Technological University)
Anastasios Kementsietsidis (Google Research)
Andrew Crotty (Carnegie Mellon University)
Anna Fariha (Microsoft)
Anton Dignós (Free University of Bozen-Bolzano)
Antonios Deligiannakis (Technical University of Crete)
Arijit Khan (Nanyang Technological University)
Anastasios Kementsietsidis (Google Research)
Andrew Crotty (Carnegie Mellon University)
Anna Fariha (Microsoft)
Anton Dignós (Free University of Bozen-Bolzano)
Antonios Deligiannakis (Technical University of Crete)
Arijit Khan (Nanyang Technological University)
LETTER FROM THE EDITORS IN CHIEF

It is our pleasure to present the eleventh issue of PVLDB (Proceedings of the VLDB) Volume 16. This is the last issue that includes research papers that will be presented at VLDB 2023 in Vancouver.

PVLDB presents original research papers on a broad range of topics related to all aspects of data and information management, spanning from theoretical foundations, system architectures, models and techniques, to novel applications as well as large-scale deployment and evaluation. There are four equally important categories of papers in the research track: (a) regular research, (b) scalable data science, (c) experiment, analysis & benchmark, and (d) vision. Each paper is evaluated by at least three reviewers and an Associate Editor, who summarizes in a meta-review all reviews and the results of a three-week discussion phase during which the reviewers exchange their viewpoints and converge to a joint decision.

This issue, the biggest so far in this volume, includes 60 papers, spanning the topics of: Data Mining and Analytics; Data Privacy and Security; Database Engines; Database Performance and Manageability; Distributed Database Systems; Graph and Network Data; Information Integration and Data Quality; Languages; Machine Learning, AI, and Databases; Novel Database Architectures; Provenance and Workflows; Specialized and Domain-Specific Data Management; Text and Semi-Structured Data; and User Interfaces (these cover all top-level topics of interest defined by our Call for Contributions). The most popular topics in this issue are:

- Database Engines (23 papers);
- Machine Learning, AI, and Databases (19 papers);
- Graph and Network Data (10 papers).

The breakdown according to the paper categories is as follows:

- Scalable data science (3 papers);
- Experiment, analysis & benchmark (8 papers);
- Vision (2 papers);
- Regular research papers (47 papers).

Out of the 60 papers, 53 were accepted after revision, and 7 were accepted after revision plus shepherding.

This issue is the result of all the work put in by the authors as well as the great commitment and effort of our associate editors and reviewers as well as our proceedings chairs.

Georgia Koutrika and Jun Yang
Editors-in-Chief of PVLDB Vol. 16
Program Chairs for VLDB 2023