Proceedings of the VLDB Endowment

Volume 14, No. 11 – July 2021

Editors in Chief:

Xin Luna Dong and Felix Naumann

Associate Editors:

Alon Halevy, Anastasia Ailamaki, Angela Bonifati, Arun Kumar, Ashraf Aboulnaga, Eugene Wu, Floris Geerts, Graham Cormode, Jeffrey Xu Yu, Jiannan Wang, Jingren Zhou, Jorge Arnulfo Quiané Ruiz, Juliana Freire, Jun Yang, Martin Theobald, Nesime Tatbul, Paolo Papotti, Rainer Gemulla, Stefan Manegold, Stratos Idreos, Surajit Chaudhuri, Xuemin Lin, Yi Chen, Yufei Tao, Zachary Ives, Zhifeng Bao

Publication Editors:

Thorsten Papenbrock and Hannes Mühleisen

PVLDB – Proceedings of the VLDB Endowment

Volume 14, No. 11, July 2021.

All papers published in this issue will be presented at the 47th International Conference on Very Large Data Bases, Copenhagen, Denmark, 2021.

Copyright 2021 VLDB Endowment

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org.

Volume 14, Number 11, July 2021

Pages i – vii and 1922 - 2654

ISSN 2150-8097

Available at: http://www.pvldb.org and https://dl.acm.org/journal/pvldb

TABLE OF CONTENTS

Front Matter

Copyright Notice i
Table of Contents ii
PVLDB Organization and Review Board – Vol. 14 iv
Editorial vii

Research Papers

GraphMineSuite: Enabling High-Performance and Programmable Graph Mining Algorithms with Set Algebra 1922
Maciej Besta, Zur Vonarburg-shmaria, Yannick Schaffner, Leonardo Schwarz, Grzegorz Kwasniewski, Lukas Gianinazzi, Jakub Beranek, Kacper Janda, Tobias Holenstein, Sebastian Leisinger, Peter Tatkowski, Esref Özdemir, Adrian Balla, Marcin Copik, Philipp Lindenberger, Marek Konieczny, Onur Mutlu, Torsten Hoefler

PATSQL: Efficient Synthesis of SQL Queries from Example Tables with Quick Inference of Projected Columns 1937
Keita Takenouchi, Takashi Ishio, Joji Okada, Yuji Sakata

Fauce: Fast and Accurate Deep Ensembles with Uncertainty for Cardinality Estimation 1950
Jie Liu, Wenqian Dong, Dong Li, Qingqing Zhou

A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search 1964
Mengzhao Wang, Xiaoliang Xu, Qiang Yue, Yuxiang Wang

Towards Plug-and-Play Visual Graph Query Interfaces: Data-driven Canned Pattern Selection for Large Networks 1979
Zifeng Yuan, Huey Eng Chua, Sourav S Bhowmick, Zekun Ye, Wook-shin Han, Byron Choi

ThunderRW: An In-Memory Graph Random Walk Engine 1992
Shixuan Sun, Yuhang Chen, Shengliang Lu, Bingsheng He, Yuchen Li

Butterfly-Core Community Search over Labeled Graphs 2006
Zheng Dong, Xin Huang, Guorui Yuan, Hengshu Zhu, Hui Xiong

Flow-Loss: Learning Cardinality Estimates That Matter 2019
Parimarjan Negi, Ryan C Marcus, Andreas Kipf, Hongzi Mao, Nesime Tatbul, Tim Kraska, Mohammad Alizadeh

On Querying Historical K-Cores 2033
Michael R Yu, Dong Wen, Lu Qin, Ying Zhang, Wenjie Zhang, Xuemin Lin

Frequency Estimation under Local Differential Privacy 2046
Graham Cormode, Sam Maddock, Carsten Maple

Doing More with Less: Characterizing Dataset Downsampling for AutoML 2059
Fatjon Zogaj, Jose P Cambronero Sanchez, Martin Rinard, Jürgen Cito

LES3: Learning-based exact set similarity search 2073
Yifan Li, Xiaohui Yu, Nick Koudas

Large Graph Convolutional Network Training with GPU-Oriented Data Communication Architecture 2087
Seung Won Min, Kun Wu, Sitao Huang, Mert Hidayetoglu, Jinjun Xiong, Eiman Ebrahimi, Deming Chen, Wen-mei Hwu

FlexPushdownDB: Hybrid Pushdown and Caching in a Cloud DBMS 2101
Yifei Yang, Matt Youill, Matthew Woicik, Yizhou Liu, Xiangyao Yu, Marco Serafini, Ashraf Aboulnaga, Michael Stonebraker

Approximating Median Absolute Deviation with Bounded Error 2114
Zhiwei Chen, Shaoxu Song, Ziheng Wei, Jingyun Fang, Jiang Long

An Experimental Evaluation and Guideline for Path Finding in Weighted Dynamic Network 2127
Mengxuan Zhang, Lei Li, Xiaofang Zhou

Robustness against Read Committed for Transaction Templates 2141
Brecht Vandevoort, Bas Ketsman, Christoph Koch, Frank Neven

LANCET: Labeling Complex Data at Scale 2154
Huayi Zhang, Lei Cao, Samuel Madden, Elke A Rundensteiner

VolcanoML: Speeding up End-to-End AutoML via Scalable Search Space Decomposition 2167
Yang Li, Yu Shen, Wentao Zhang, Jiawei Jiang, Yaliang Li, Bolin Ding, Jingren Zhou, Zhi Yang, Wentao Wu, Ce Zhang, Bin Cui

A Queueing-Theoretic Framework for Vehicle Dispatching in Dynamic Car-Hailing 2177
Peng Cheng, Jiabao Jin, Lei Chen, Xuemin Lin, Libin Zheng

Data Synthesis via Differentially Private Markov Random Field 2190
Kuntai Cai, Xiaoyu Lei, Jianxin Wei, Xiaokui Xiao

Scaling Replicated State Machines with Compartmentalization 2203
Michael J Whittaker, Ailidani Ailidani, Aleksey Charapko, Murat Demirbas, Neil Giridharan, Joseph M Hellerstein, Heidi Howard, Ion Stoica, Adriana Szekeres

Constructing and Analyzing the LSM Compaction Design Space 2216
Subhadeep Sarkar, Dimitris Staratzis, Zichen Zhu, Manos Athanassoulis

ByShard: Sharding in a Byzantine Environment 2230
Jelle Hellings, Mohammad Sadoghi

SetSketch: Filling the Gap between MinHash and HyperLogLog 2244
Otmar Ertl

CGM: An Enhanced Mechanism for Streaming Data Collectionwith Local Differential Privacy 2258
Ergute Bao, Yin Yang, Xiaokui Xiao, Bolin Ding

Errata for "Teseo and the Analysis of Structural Dynamic Graph" 2271
Dean De Leo, Per Fuchs, Peter Boncz

QARTA: An ML-based System for Accurate Map Services 2273
Mashaal Musleh, Sofiane Abbar, Rade Stanojevic, Mohamed Mokbel

Real-World Trajectory Sharing with Local Differential Privacy 2283
Teddy Cunningham, Graham Cormode, Hakan Ferhatosmanoglu, Divesh Srivastava

PolyFrame: A Retargetable Query-based Approach to Scaling Dataframes 2296
Phanwadee Sinthong, Michael Carey

Scalable Community Detection via Parallel Correlation Clustering 2305
Jessica Shi, Laxman Dhulipala, David Eisenstat, Jakub Łącki, Vahab Mirrokni

SlimChain: Scaling Blockchain Transactions through Off-Chain Storage and Parallel Processing ... 2314
Cheng Xu, Ce Zhang, Jianliang Xu, Jian Pei

Towards an Optimized GROUP BY Abstraction for Large-Scale Machine Learning 2327
Side Li, Arun Kumar

Accelerating Approximate Aggregation Queries with Expensive Predicates 2341
Daniel Kang, John Guibas, Peter D Bailis, Tatsunori Hashimoto, Yi Sun, Matei Zaharia

A four-dimensional Analysis of Partitioned Approximate Filters 2355
Tobias Schmidt, Maximilian Bandle, Jana Giceva

SKT: A One-Pass Multi-Sketch Data Analytics Accelerator 2369
Monica Chiosa, Thomas Preußer, Gustavo Alonso

A Practical Approach to Groupjoin and Nested Aggregates 2383
Philipp Fent, Thomas Neumann

Robust Voice Querying with MUVE: Optimally Visualizing Results of Phonetically Similar Queries. 2397
Ziyun Wei, Immanuel Trummer, Connor Anderson

CHEF: A Cheap and Fast Pipeline for Iteratively Cleaning Label Uncertainties 2410
Yinjun Wu, James Weimer, Susan B Davidson

COMPARE: Accelerating Groupwise Comparison in Relational Databases for Data Analytics 2419
Tarique Siddiqui, Surajit Chaudhuri, Vivek Narasayya

Crystal: A Unified Cache Storage System for Analytical Databases 2432
Dominik Durner, Badrish Chandramouli, Yinan Li

The Smallest Extraction Problem 2445
Valerio Cetorelli, Paolo Atzeni, Valter Crescenzi, Franco Milicchio

Deep Learning for Blocking in Entity Matching: A Design Space Exploration 2459
Saravanan Thirumuruganathan, Han Li, Nan Tang, Mourad Ouzzani, Yash Govind, Derek Paulsen, Glenn M Fung, Anhai Doan

Grain: Improving Data Efficiency of Graph Neural Networks via Diversified Influence Maximization 2473
Wentao Zhang, Zhi Yang, Yexin Wang, Yu Shen, Yang Li, Liang Wang, Bin Cui

Database Technology for the Masses: Sub-Operators as First-Class Entities 2483
Maximilian Bandle, Jana Giceva

Columnar Storage and List-based Processing for Graph Database Management Systems 2491
Pranjal Gupta, Amine Mhedhbi, Semih Salihoglu

Phoebe: A Learning-based Checkpoint Optimizer 2505
Yiwen Zhu, Matteo Interlandi, Abhishek Roy, Krishnadhan Das, Hiren Patel, Malay Bag, Hitesh Sharma, Alekh Jindal

Tailoring Data Source Distributions for Fairness-aware Data Integration 2519
Fatemeh Nargesian, Abolfazl Asudeh, H. V. Jagadish

Missing Value Imputation on Multidimensional Time Series 2533
Parikshit Bansal, Prathamesh Deshpande, Sunita Sarawagi

Horizon: Scalable Dependency-driven Data Cleaning 2546
El Kindi Rezig, Mourad Ouzzani, Walid G. Aref, Ahmed Elmagarmid, Ahmed Mahmood, Michael Stonebraker

Declarative Data Serving: The Future of Machine Learning Inference on the Edge 2555
Ted Shaowang, Nilesh Jain, Dennis Matthews, Sanjay Krishnan

Auto-Pipeline: Synthesize Data Pipelines By-Target Using Reinforcement Learning and Search ... 2563
Junwen Yang, Yeye He, Surajit Chaudhuri

Explaining Inference Queries with Bayesian Optimization 2576
Brandon Lockhart, Jinglin Peng, Weiyuan Wu, Jiannan Wang, Eugene Wu

Decomposed Bounded Floats for Fast Compression and Queries 2586
Chunwei Liu, Hao Jiang, John Paparrizos, Aaron J Elmore

Beyond Equi-joins: Ranking, Enumeration and Factorization 2599
Nikolaos Tziavelis, Wolfgang Gatterbauer, Mirek Riedewald

Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series 2613
Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, Nesime Tatbul

Progressive Compressed Records: Taking a Byte out of Deep Learning Data 2627
Michael Kuchnik, George Amvrosiadis, Virginia Smith

TQEL: Framework for Query-Driven Linking of Top-K Entities in Social Media Blogs 2642
Abdulrahman Alsaudi, Yasser Altowim, Sharad Mehrotra, Yaming Yu

PVLDB ORGANIZATION AND REVIEW BOARD - Vol. 14

Editors in Chief of PVLDB
Xin Luna Dong (Amazon)
Felix Naumann (HPI, University of Potsdam)

Associate Editors of PVLDB
Ashraf Aboulnaga (Qatar Computing Research Institute, Hamad Bin Khalifa University)
Anastasia Ailamaki (EPFL)
Zhifeng Bao (RMIT University)
Angela Bonifati (Lyon 1 University)
Surajit Chaudhuri (Microsoft Research)
Yi Chen (New Jersey Institute of Technology)
Graham Cormode (University of Warwick)
Juliana Freire (New York University)
Floris Geerts (University of Antwerp)
Rainer Gemulla (University of Mannheim)
Alon Halevy (Facebook)
Stratos Idreos (Harvard University)
Zachary Ives (University of Pennsylvania)
Arun Kumar (UC San Diego)
Xuemin Lin (University of New South Wales)
Stefan Manegold (CWI, Leiden University)
Paolo Papotti (Eurecom)
Jorge Arnulfo Quiané Ruiz (Technical University of Berlin)
Yufei Tao (Chinese University of Hong Kong)
Nesime Tatbul (Intel Labs and MIT)
Martin Theobald (Université du Luxembourg)
Jiannan Wang (Simon Fraser University)
Eugene Wu (Columbia University)
Jun Yang (Duke University)
Jeffrey Xu Yu (The Chinese University of Hong Kong)
Jingren Zhou (Alibaba)

Publication Editors
Thorsten Papenbrock (HPI, University of Potsdam)
Hannes Mühleisen (CWI)

PVLDB Managing Editor
Wolfgang Lehner (Dresden University of Technology)

PVLDB Advisory Committee
Divesh Srivastava (AT&T Labs-Research)
M. Tamer Özsu (University of Waterloo)
Juliana Freire (New York University)
Xin Luna Dong (Amazon)
Peter Boncz (CWI)
Lei Chen (Hong Kong University of Science and Technology)
Graham Cormode (University of Warwick)
Xiaofang Zhou (University of Queensland)
Magdalena Balazinska (University of Washington)
Fatma Ozcan (IBM Almaden)
Felix Naumann (HPI, University of Potsdam)
Peter Triantafillou (University of Warwick)

Review Board

Abolfazl Asudeh (University of Illinois)
Ahmed Eldawy (University of California, Riverside)
Alan Fekete (University of Sydney)
Alekh Jindal (Microsoft)
Alex Ratner (University of Washington)
Altigran da Silva (Universidade Federal do Amazonas)
Anthony Tung (National University of Singapore)
Antonios Deligiannakis (Technical University of Crete)
Arijit Khan Nanyang (Technological University, Singapore)
Arnau Prat (Sparsity Technologies)
Ashwin Machanavajjhala (Duke University)
Asterios Katsifodimos (Technical University of Delft)
Avrilia Floratou (Microsoft)
Babak Salimi (University of Washington)
Badrish Chandramouli (Microsoft Research)
Beng Chin Ooi (National University of Singapore)
Bin Yang (Aalborg University)
Boris Glavic (Illinois Institute of Technology)
Byron Choi (Hong Kong Baptist University)
Carlos Scheidegger (University of Arizona)
Carsten Binnig (Technical University of Darmstadt)
Ce Zhang (ETH Zurich)
Chengfei Liu (Swinburne University of Technology)
Chengkai Li (University of Texas at Arlington)
Chris Jermaine (Rice University)
Christian Bizer (University of Mannheim)
Cong Yu (Google)
Daisy Zhe Wang (University of Florida)
Danica Porobic (Oracle)
Davide Mottin (Aarhus University)
Dimitris Papadias (Hong Kong University of Science and Technology)
Dong Deng (Rutgers University)
Eric Lo (Chinese University of Hong Kong)
Essam Mansour (Concordia University)
Fatma Ozcan (IBM Research)
Flip Korn (Google)
Florin Rusu (University of California, Merced)
Fotis Psallidas (Microsoft)
Francesco Bonchi (ISI Foundation)
Gao Cong (Nanyang Technological University)
George Fletcher (Technical University of Eindhoven)
Georgia Koutrika (Athena Research Center)
Hao Wei (Amazon)
Heiko Mueller (New York University)
Hong Cheng (Chinese University of Hong Kong)
Hongzhi Wang (Harbin Institute of Technology)
Hung Ngo (RelationalAI)
Immanuel Trummer (Cornell University)
Ingo Müller (ETH Zürich)
Jana Giceva (Technical University of Munich)
Jennie Rogers (Northwestern University)
Jeong-Hyon Hwang (University at Albany, State University of New York)
Jiaheng Lu (University of Helsinki)
Jianliang Xu (Hong Kong Baptist University)
Jianxin Li (Deakin University)
Jignesh Patel (University of Wisconsin)
Johann Gamper (Free University of Bozen-Bolzano)
Johannes Gehrke (Microsoft)
Jonas Traub (Technical University of Berlin)
Joy Arulraj (Georgia Tech)
Ju Fan (Renmin University of China)
K. Selçuk Candan (Arizona State University)
Kai Zeng (Alibaba)
Katja Hose (Aalborg University)
Ken Salem (University of Waterloo)
Kenneth A. Ross (Columbia University)
Khuzaima Daudjee (University of Waterloo)
Konstantinos Karanasos (Microsoft)
Laurel Orr (Stanford University)
Lei Chen (Hong Kong University of Science and Technology)
Lei Zou (Peking University)
Li Xiong (Emory University)
Lu Chen (Aalborg University)
Lu Qin (University of Technology Sydney)
Manasi Vartak (Verta)
Manos Athanassoulis (Boston University)
Manos Karpathiotakis (Facebook)
Marco Serafini (University of Massachusetts Amherst)
Marcos Antonio Vaz Salles (University of Copenhagen)
Mark Callaghan (MongoDB)
Markus Weimer (Microsoft)
Matei Zaharia (Stanford University, Databricks)
Matteo Interlandi (Microsoft)
Matthaios Olma (Microsoft Research)
Meihui Zhang Beijing (Institute of Technology)
Miao Qiao (University of Auckland)
Michael H. Böhlen (University of Zurich)
Michael Cafarella (University of Michigan)
Mirek Riedewald (Northeastern University)
Mohamed Mokbel (Qatar Computing Research Institute)
Mohamed Sarwat (Arizona State University)
Mohammad Sadoghi (University of California, Davis)
Mourad Ouzzani (Qatar Computing Research Institute, Hamad Bin Khalifa University)
Muhammad Aamir Cheema (Monash University)
Murat Demirbas (University at Buffalo, SUNY)
Nan Tang (Qatar Computing Research Institute, Hamad Bin Khalifa University)
Nick Koudas (University of Toronto)
Nikolaus Augsten (University of Salzburg)
Norman May (SAP)
Norman Paton (University of Manchester)
Odysseas Papapetrou (Technical University of Eindhoven)
Oliver A. Kennedy (University at Buffalo, SUNY)
Paolo Merialdo (Roma Tre University)
Paraschos Koutris (University of Wisconsin – Madison)
Peter Boncz (Centrum Wiskunde & Informatica)
Qin Zhang Indiana (University Bloomington)
Raja Appuswamy (Eurecom)
Ralf Schenkel (University of Trier)

Raul Castro Fernandez (University of Chicago)
Raymond Chi-Wing Wong (Hong Kong University of Science and Technology)
Reynold Cheng (The University of Hong Kong)
Reza Akbarinia (INRIA)
Ruoming Jin (Kent State University)
Ryan Johnson (Amazon Web Services)
S. Sudarshan (IIT Bombay)
Sanjay Krishnan (University of Chicago)
Saravanan Thirumuruganathan (Qatar Computing Research Institute, Hamad Bin Khalifa University)
Sebastian Schelter (University of Amsterdam)
Semih Salihoglu (University of Waterloo)
Senjuti Basu Roy (New Jersey Institute of Technology)
Shaoxu Song (Tsinghua University)
Shimin Chen (Chinese Academy of Sciences)
Sibo Wang (The Chinese University of Hong Kong)
Silu Huang (Microsoft Research)
Spyros Blanas (Ohio State University)
Srikanth Kandula (Microsoft Research)
Steffen Zeuch (German Research Centre for Artificial Intelligence - DFKI)
Stijn Vansummeren (Université libre de Bruxelles)
Sudeepa Roy (Duke University)
Sudip Roy (Google)
Tamer Özsu (University of Waterloo)
Themis Palpanas (University of Paris, French University Institute - IUF)
Tianzheng Wang (Simon Fraser University)
Tingjian Ge (University of Massachusetts, Lowell)
Thomas Heinis (Imperial College)
Thomas Neumann (Technical University of Munich)
Toon Calders (Universiteit Antwerpen)
Umar Farooq Minhas (Microsoft Research)
Viktor Leis (Friedrich Schiller University Jena)
Walid Aref (Purdue University)
Wei-Shinn Ku (Auburn University)
Weiren Yu (University of Warwick)
Wendy Hui Wang (Stevens Institute of Technology)
Wenjie Zhang (University of New South Wales)
Wolfgang Gatterbauer (Northeastern University)
Xi He (University of Waterloo)
Xiang Zhao (National University of Defence Technology)
Xiangyao Yu (University of Wisconsin – Madison)
Xiaokui Xiao (National University of Singapore)
Xiaolan Wang (Megagon Labs)
Xin Cao (University of New South Wales)
Xu Chu (Georgia Tech)
Yannis Velegrakis (Utrecht University)
Ye Yuan (Beijing Institute of Technology)
Yeye He (Microsoft Research)
Ying Zhang (University of Technology Sydney)
Yinghui Wu (Case Western Reserve University)
Yongjoo Park (University of Illinois at Urbana-Champaign)
Yongxin Tong (Beihang University)
Yu Yang (City University of Hong Kong)
Yuchen Li (Singapore Management University)
Yudian Zheng (Twitter)
Yunjun Gao (Zhejiang University)
Zechao Shang (University of Chicago)
Zhenjie Zhang (Singapore R&D, Yitu Technology Ltd.)
Zhewei Wei (Renmin University of China)
Ziawasch Abedjan (Technical University of Berlin)
Zoi Kaoudi (Technical University of Berlin)

EDITORIAL

We are delighted to present to you the 11th issue of the Proceedings of the VLDB Endowment (PVLDB), Volume 14. PVLDB publishes papers accepted through a journal-style reviewing process from papers submitted on a year-round monthly submission schedule. All PVLDB papers published in this volume will be presented at the VLDB 2021 Conference, to be held in Copenhagen, Denmark, from August 16 to August 20, 2021.

This issue of PVLDB showcases fifty-eight excellent papers. These papers span a broad spectrum of topics, demonstrating the healthy diversity of data-centric issues our community is investigating. Notable topics in this issue address learned selectivity estimation, approximate nearest neighbor search, learned entity matching, scaling replicated state machines, scaling dataframes, scalable AutoML, fairness-preserving data integration, and privacy preserving statistics. We thank all the authors for making high-quality submissions and working with the PC to improve them!

In addition, as the associate editors for the Experiment, Analysis and Benchmark paper category at PVLDB this year, we are delighted about the breadth and excellence of the submissions in this category overall. Papers were clustered under three categories: 1) in-depth experimental analyses of particular systems or algorithms, 2) design space explorations that experimentally investigate nearly all possible dimensions of a problem, and for the first time this year 3) new benchmarks that enable a deeper comparison of systems in a mature or maturing application area. The benchmarks category is new for 2021, marking an expansion of this track to further include types of papers that can have broad impact both in how methods are developed and in applications. The high-order bit of what makes an excellent paper in this category (and what our review and revision process was focused on) is the ability to provide new insights useful to the research community, through an in-depth experimentation process that exceeds that of the original papers presenting the given techniques. Similarly, for papers that proposed new benchmarks, we were focused on the ability of the benchmark to generate new and practical insights.

We hope that you enjoy our selection of papers in this issue, and those from the Experiment, Analysis and Benchmark category overall. We thank all authors and all reviewers for their exceptional work and discussions, in making the best selections to present to our community - and we hope that you will continue to submit your best work to PVLDB's future issues!

Stratos Idreos and Zack Ives