# TABLE OF CONTENTS

**Front Matter**

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

**Research Papers**

Benchmarking Learned Indexes .................................................................................................................. 1
*Ryan Marcus, Andreas Kipf, Alexander Van Renen, Mihail Stoian, Sanchit Misra, Alfons Kemper, Thomas Neumann, Tim Kraska*

Tempura: A General Cost-Based Optimizer Framework for Incremental Data Processing ............. 14
*Zuozhi Wang, Kai Zeng, Botong Huang, Wei Chen, Xiaozong Cui, Bo Wang, Ji Liu, Liya Fan, Dachuan Qu, Zhenyu Hou, Tao Guan, Chen Li, Jingren Zhou*

Inspector Gadget: A Data Programming-based Labeling System for Industrial Images .............. 28
*Geon Heo, Yuji Roh, Seonghyeon Hwang, Dayun Lee, Steven Whang*

Scaling Attributed Network Embedding to Massive Graphs ................................................................. 37
*Renchi Yang, Jieming Shi, Xiaokui Xiao, Yin Yang, Juncheng Liu, Sourav S. Bhowmick*

Deep Entity Matching with Pre-Trained Language Models ................................................................. 50
*Yuliang Li, Jinfeng Li, Yoshihiko Suhara, Anhai Doan, Wang-Chiew Tan*

NeuroCard: One Cardinality Estimator for All Tables ........................................................................... 61
*Zongheng Yang, Amog Kamsetty, Sifei Luan, Eric Liang, Yan Duan, Peter Chen, Ion Stoica*
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)
Suajit Chaudhuri (Microsoft Research)
Yi Chen (New Jersey Institute of Technology)
Graham Cormode (University of Warwick)
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)
LETTER FROM THE EDITORS IN CHIEF

We are very pleased to present the first issue of PVLDB’s Volume 14. The Proceedings of the VLDB present the latest research in the area of database and information system technology. Together with expert boards of associate editors and reviewers, submissions are carefully peer-reviewed, often entering a revision phase, then published in the journal and ultimately presented at the following VLDB conference. We are very grateful for all the great colleagues who contribute to the ongoing success of PVLDB. Before introducing the individual papers in this first issue, we would like to report on some changes for PVLDB that were implemented for this volume and of which we hope they improve the scope of the journal, the quality and rapid dissemination of the research, the quality of the reviewing process, and finally the availability and reproducibility of database research.

A new paper category, Scalable Data Science (SDS), seeks to accept papers that describe the design, implementation, experience, and evaluation of solutions and systems for practical data science and data engineering tasks on large-scale data. These papers are limited to 8 pages of content and do not necessarily propose new breakthrough algorithms or models, but emphasize solutions that either solve or advance the understanding of issues related to data science technologies in the real world. In the first months, about 15% of all submissions were in this SDS category. Another category change is the extension of the traditional Experiments and Analysis (E&A) category to include Benchmarks (EA&B). We recognize that benchmarks are drivers of research, providing experimental settings and a common evaluation for relevant problems. Designing, testing and publishing such benchmarks are important research contributions. Among the six papers of this first edition are one SDS and one EA&B contribution.

PVLDB strives to give high quality and constructive feedback in the form of reviews and meta-reviews. The latter summarize the reviews and the results of the three-week discussion phase in which reviewers exchange their view of the paper and converge to a joint decision. In the past, positive decisions were restricted to either a direct “accept” or a “major revision”. To account for submissions that are already close to an accept decision and for which reviewers had only smaller requests, we have introduced a “minor revision”. Among the papers of this issue, two had originally received a minor revision decision and were ultimately accepted.

Finally, our community has recognized that reproducibility, repeatability, and replicability are very important factors to advance database research. The ability to re-use code, data and experimental settings and thus compare one’s own result to previous work allows independent validation of the previous results and accelerates the advancement of state of the art. For PVLDB’s volume 14 we have introduced several measures to encourage authors to make available their research artifacts. First, we allow authors to submit supplemental material, such as data and code, and explain those files in a submission form question. In turn, reviewers are asked to answer a review question about the availability and potential reproducibility of the presented research. Finally, we have instituted an optional new metadata tag on the front page of accepted papers, namely an “Artifacts available” tag along with an author-supplied stable URL. PVLDB’s reproducibility chairs verify the availability of research artifacts, yet without performing reproducibility checks.

For this issue, the review board has selected six contributions, addressing both core database technology and work about non-relational data management. Interestingly, five of these contributions make direct use of machine learning technologies. Marcus et al. present an indexing benchmark specifically designed to evaluate and compare the novel so-called learned indexes, which train a model of the underlying data instead of relying on a pre-determined data structure. Wang et al. also address a traditional database problem, namely that of query plan cost estimation, but design a cost model for incremental data processing tasks as they appear for instance in stream processing systems. Yang et al. also apply machine learning to improve core database performance, in this case learning join cardinalities - an important ingredient in cost models. Heo et al. report on their solution to employ weak supervision in the form of labeling functions to annotate industrial images to identify and locate defects, such as scratches. Yang et al. address the important problem of scalability for the complex task of computing attributed network embeddings without compromising the quality of the embedding. Finally, Li et al. address the well-known entity matching task by applying language models, for instance to the sometimes-complex names of products.

All papers will be presented at the 2021 Conference on Very Large Databases (VLDB’21) in Copenhagen. We hope you enjoy reading and look forward to seeing you there.

Xin Luna Dong and Felix Naumann
Editors-in-Chief of PVLDB Volume 14
Program Chairs for VLDB 2021