

VLDB2012

38th International Conference on Very Large Data Bases, Istanbul, Turkey



Proceedings of the VLDB Endowment

Volume 5, No. 10 – June 2012

**Proceedings of the 38th International Conference on
Very Large Data Bases, Istanbul, Turkey**

Editor-in-Chief:

Z. Meral Özsoyoğlu

Associate Editors - Research Track:

Uğur Çetintemel, Nilesh Dalvi, Hank Korth, Anthony Tung

Associate Editors - Experiments and Analysis Track:

Gustavo Alonso, Juliana Freire

Proceedings Editors:

Ahmet Saçan, Nesime Tatbul

PVLDB – Proceedings of the VLDB Endowment

Volume 5, No. 10, June 2012.

The 38th International Conference on Very Large Data Bases, Istanbul, Turkey.

Copyright 2012 VLDB Endowment

Permission to make digital or hard copies of portions of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than VLDB Endowment must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists requires prior specific permission and/or a fee. Request permission to republish from PVLDB under email: info@vldb.org.

Volume 5, Number 10: VLDB 2012

Pages i - viii and 908 - 1087

ISSN 2150-8097, June 2012.

Additional copies only online at: portal.acm.org, arxiv.org/corr, and www.vldb.org

TABLE OF CONTENTS

Front Matter

Copyright Notice	ii
Table of Contents	iii
PVLDB Review Board	v

Letters

Letter from the Associate Editors	viii
..... <i>Ahmet Saçan, Nesime Tatbul</i>	

Research Track Papers

Answering Table Queries on the Web using Column Keywords	908
..... <i>Rakesh Pimplikar, Sunita Sarawagi</i>	
Efficient Verification of Web-Content Searching Through Authenticated Web Crawlers	920
<i>Michael T. Goodrich, Duy Nguyen, Olga Ohrimenko, Charalampos Papamanthou, Roberto Tamassia, Nikos Triandopoulos, Cristina Videira Lopes</i>	
SODA: Generating SQL for Business Users	932
..... <i>Lukas Blunschi, Claudio Jossen, Donald Kossman, Magdalini Mori, Kurt Stockinger</i>	
Privacy Preservation by Disassociation	944
..... <i>Manolis Terrovitis, John Liagouris, Nikos Mamoulis, Spiros Skiadopoulos</i>	
Supercharging Recommender Systems using Taxonomies for Learning User Purchase Behavior .	956
<i>. Bhargav Kanagal, Amr Ahmed, Sandeep Pandey, Vanja Josifovski, Jeff Yuan, Lluís Garcia-Pueyo</i>	
DBToaster: Higher-order Delta Processing for Dynamic, Frequently Fresh Views	968
..... <i>Yanif Ahmad, Oliver Kennedy, Christoph Koch, Milos Nikolic</i>	
Real Time Discovery of Dense Clusters in Highly Dynamic Graphs: Identifying Real World Events in Highly Dynamic Environments	980
..... <i>Manoj K Agarwal, Krithi Ramamritham, Manish Bhide</i>	
Sketch-based Querying of Distributed Sliding-Window Data Streams	992
..... <i>Odysseas Papapetrou, Minos Garofalakis, Antonios Deligiannakis</i>	
LogBase: A Scalable Log-structured Database System in the Cloud	1004
..... <i>Hoang Tam Vo, Sheng Wang, Divyakant Agrawal, Gang Chen, Beng Chin Ooi</i>	
Efficient Processing of k Nearest Neighbor Joins using MapReduce	1016
..... <i>Wei Lu, Yanyan Shen, Su Chen, Beng Chin Ooi</i>	
Early Accurate Results for Advanced Analytics on MapReduce	1028
..... <i>Nikolay Laptev, Kai Zeng, Carlo Zaniolo</i>	

CDAS: A Crowdsourcing Data Analytics System	1040
..... <i>Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, Meihui Zhang</i>	
Mining Statistically Significant Substrings using the Chi-Square Statistic	1052
..... <i>Mayank Sachan, Arnab Bhattacharya</i>	
Massively Parallel Sort-Merge Joins in Main Memory Multi-Core Database Systems	1064
..... <i>Martina-Cezara Albutiu, Alfons Kemper, Thomas Neumann</i>	
hStorage-DB: Heterogeneity-aware Data Management to Exploit the Full Capability of Hybrid Storage Systems	1076
..... <i>Tian Luo, Rubao Lee, Michael Mesnier, Feng Chen, Xiaodong Zhang</i>	

PVLDB REVIEW BOARD

VLDB 2012 General PC Chair

Z. Meral Özsoyoğlu, Case Western Reserve University, USA

VLDB 2012 Research Track PC Co-Chairs

Uğur Çetintemel, Brown University, USA

Nilesh Dalvi, Yahoo Research

Hank Korth, Lehigh University, USA

Anthony Tung, National Univ. of Singapore, Singapore

VLDB 2012 Experiments and Analysis Track PC Co-Chairs

Gustavo Alonso, ETH Zurich, Switzerland

Juliana Freire, NYU Poly, USA

VLDB 2012 Research Track Reviewers

Karl Aberer, EPFL, Switzerland

Ashraf Aboulnaga, University of Waterloo

Yanif Ahmad, Johns Hopkins University

Shivnath Babu, Duke University

Magdalena Balazinska, University of Washington

Elisa Bertino, Purdue University

Claudio Bettini, University of Milano

Sourav Bhowmick, Nayang Technological University

Angela Bonifati, Icar-CNR

Randal Burns, Johns Hopkins University

Michael Cafarella, Michigan

Kevin Chang, Univ. of Illinois at Urbana-Champaign

Gang Chen, Zhejiang University, China

Lei Chen, Hong Kong University of SaT, China

Reynold Cheng, University of Hong Kong

Jan Chomicki, University of Buffalo

Chris Clifton, Purdue University

Gao Cong, Nayang Technological University

Brian Cooper, Google, USA

Graham Cormode, AT&T

Bin Cui, Peking University

Carlo Curino, MIT

Anish Das Sharma, Yahoo

Antonios Deligiannakis, Technical University of Crete

Amr El Abbadi, UC Santa Barbara

Alan Fekete, University of Sydney

Hakan Ferhatosmanoğlu, Ohio State University

Alvaro Fernandes, University of Manchester

Johann-Christoph Freytag, Humboldt-Universität zu Berlin

Minos Garofalakis, TU Creta, Greece

Bugra Gedik, IBM Watson
 Gabriel Ghinita, Purdue University
 Goetz Graefe, HP
 Sergio Greco, University of Calabria
 Todd Green, University of California, Davis
 Le Gruenwald, University of Oklahoma
 Haryadi Gunawi, University of California, Berkeley
 Rahul Gupta, IIT Bombay
 Stavros Harizopoulos, HP
 Howard Ho, IBM
 Vagelis Hristidis, Florida International University
 Meichun Hsu, HP
 Jeong-Hyon Hwang, SUNY, Albany
 Ihab Ilyas, University of Waterloo
 Zachary Ives, University of Pennsylvania
 Dean Jacobs, SAP
 Hans-Arno Jacobsen, University of Toronto
 Christian Jensen, Aalborg University
 Chris Jermaine, Rice University
 Ricardo Jumenez-Peris, Technical University of Madrid
 Theodore Johnson, AT&T
 Vana Kalogeraki, Athens U of Econ. and Business
 Raghav Kaushik, Microsoft Research
 Eamonn Keogh, University of California, Riverside
 Martin Kersten, CWI
 Benny Kimelfeld, IBM
 George Kollios, Boston University
 Flip Korn, AT&T Labs Research
 Donald Kossmann, ETH Zurich
 Laks V.S. Lakshmanan, University of British Columbia
 Mong-Li Lee, National University of Singapore
 Wang-Chien Lee, Penn State University
 Cuiping Li, Renmin University of China
 Guoliang Li, Tsinghua University, China
 Lipyeow Lim, University of Hawaii at Manoa
 David Lomet, Microsoft Research
 Boon Thau Loo, University of Pennsylvania
 Qiong Luo, Hong Kong University of SaT, China
 Ashwin Machanavajjhala, Yahoo
 Sanjay Madria, University of Missouri-Rolla
 Stefan Manegold, CWI, Netherlands
 Sharad Mehrotra, University of California, Irvine
 Gerome Miklau, Univ. of Massachusetts, Amherst
 Bongki Moon, University of Arizona
 Kyriakos Mouratidis, Singapore Management Univ.
 Gero Muhl, University of Rostock
 Suman Nath, MSR
 Sylvia Nittel, University of Maine
 Dan Olteanu, Oxford
 M. Tamer Özsu, University of Waterloo
 Esther Pacitti, University of Montpellier
 Olga Papaemmanouil, Brandeis University
 Srinivasan Parthasarathy, The Ohio State University
 Jian Pei, Simon Fraser University
 Peter Pietzuch, Imperial College
 Lucian Popa, IBM
 Rachel Pottinger, University of British Columbia
 Sunil Prabhakar, Purdue University
 Vijayshankar Raman, IBM, USA
 Vibhar Rastogi, Yahoo
 Christopher Re, University of Wisconsin, Madison
 Ken Ross, Columbia University
 Elke Rundensteiner, Worcester Polytechnic Institute
 Maria Sapino, University of Torino
 Yücel Saygın, Sabanci University
 Kyuseok Shim, Seoul National University
 Lidan Shou, Zhejiang University, China
 Adam Silberstein, Yahoo!, USA
 Yannis Sismanis, IBM, USA
 Richard Snodgrass, University of Arizona
 Divesh Srivastava, AT&T
 Dan Suciu, University of Washington

S Sudarshan, IIT Bombay
Egemen Tanin, University of Melbourne
Yufei Tao, Chinese University of Hong Kong, China
Martin Theobald, Max Planck Institute, Germany
Frank Tompa, University of Waterloo
Stratis Viglas, University of Edinburgh
Wei Wang, University of North Carolina

Gerhard Weikum, Max Planck Institute, Germany
Dong Xin, Google
Jiong Yang, Case Western Reserve University
Jun Yang, Duke
Ge Yu, Northeastern University, China
Aoying Zhou, East China Normal University, China

VLDB 2012 Experiments and Analysis Track Reviewers

Cristiana Amza, U. Toronto, Canada
Phil Bernstein, Microsoft Research, USA
Philippe Bonnet, IT U. of Copenhagen, Denmark
Michael Carey, UC Irvine, USA
Jens Dittrich, Uni. Saarland, Germany
Mary Fernandez, AT&T Research, USA
Bettina Kemme, McGill Uni., Canada
Whang Kyu-Young, KAIST, Korea
Samuel Madden, MIT, USA

Ioana Manolescu, INRIA, France
Dimitris Papadias, Hong Kong University of SaT, China
Hamid Pirahesh, IBM Almaden, USA
Maya Ramanath, IIT Delhi, India
Uwe Roehm, University of Sydney, Australia
Jens Teubner, ETH Zurich, Switzerland
Peter Triantafillou, U. Patras, Greece
Cong Yu, Google, USA

PVLDB Information Director

Gerald Weber, University of Auckland, New Zealand

VLDB 2012 Proceedings Co-Chairs

Ahmet Saçan, Drexel University, USA
Nesime Tatbul, ETH Zurich, Switzerland

Advisory Committee

Philip Bernstein, Michael Boehlen, Peter Buneman, Susan Davidson, Z. Meral Özsoyoğlu, S. Sudarshan, Gerhard Weikum

LETTER FROM THE ASSOCIATE EDITORS

It is our pleasure to present the tenth issue of Proceedings of the VLDB Endowment (PVLDB) Volume 5. This issue contains fifteen papers, all from the Research Track. These papers cover a wide range of topics, including web and data warehouse search, data anonymization, recommender systems, incremental view maintenance, distributed stream processing, cloud databases, MapReduce, crowdsourcing, sequence mining, modern hardware, and heterogeneous storage management. These and other papers appearing in PVLDB Volume 5 will be presented at the VLDB 2012 Conference, to be held in Istanbul, Turkey, August 27-31, 2012. All of these papers were reviewed in a journal-style review and revision process with monthly year-round submission cycles, with the exception that the revision option is replaced by author feedback for the last two submission cycles to ensure timely inclusion in Volume 5.

As Proceedings Chairs, we have enjoyed perusing the excellent papers accepted to VLDB 2012. We thank the authors for their high-quality work and for their readiness and haste to incorporate our suggested formatting and cosmetic changes into their papers before publication in PVLDB. We also thank the Review Board (program committee) for ensuring scientific and technical quality of the papers before they reached us. It has been a pleasure to work with an outstanding set of editorial colleagues, the Editor-in-Chief (program committee chair) and the Associate Editors (PC co-chairs), who have orchestrated the review process and have guided us in executing an innovative publication model.

At the time of publication of this issue of PVLDB, an Advance Program for VLDB 2012 has been announced, previewing the papers to be published in the remaining issues of Volume 5 and showcasing an exciting set of keynote speakers, tutorials, panels, demos, and industrial papers.

It has no doubt been a busy year for everyone involved in VLDB 2012 and we look forward to seeing you in Istanbul to enjoy an outstanding conference together!

Ahmet Saçan, Drexel University, USA

Nesime Tatbul, ETH Zurich, Switzerland

Associate Editors, Proceedings of the VLDB Endowment, Volume 5

Proceedings Co-Chairs, VLDB 2012