



## Proceedings of the VLDB Endowment

Volume 4, No. 12 – Industrial, Applications and Experience Track,  
Demonstrations, Challenges & Vision Track, Tutorials and Panels  
**Proceedings of the 37th International Conference on  
Very Large Data Bases, Seattle, WA**

Editor-in-Chief:

**H. V. Jagadish**

Guest Editors:

**José Blakeley, Joseph M. Hellerstein, Nick Koudas, Wolfgang Lehner, Sunita Sarawagi, Uwe Röhm**

PVLDB – Proceedings of the VLDB Endowment

Volume 4, No. 12, August 2011.

The 37th International Conference on Very Large Data Bases, Seattle, WA.

## **Copyright 2011 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](mailto:info@vldb.org).

Volume 4, Number 12:  
VLDB2011 Industrial Track, Demonstrations, Challenges and Vision Track,  
Tutorials and Panels  
Pages ii – viii and 1213 – 1515  
ISSN 2150-8097, August 2011.

Additional copies only online at: [portal.acm.org](http://portal.acm.org), [arxiv.org/corr](http://arxiv.org/corr), and [www.vldb.org](http://www.vldb.org)

## TABLE OF CONTENTS

### Front Matter

Copyright Notice .....	ii
Table of Contents .....	iii
PVLDB Program Committees .....	vii
Letter from the Industrial Track Chairs..... <i>Berthold Reinwald, Phil Bohannon</i>	ix

### Industrial, Applications, and Experience Track Papers

Evaluation Strategies for Top-k Queries over Memory-Resident Inverted Indexes ..... 1213 ..... <i>Marcus Fontoura, Vanja Josifovski, Jinhui Liu, Srihari Venkatesan, Xiangfei Zhu, Jason Zien</i>	1213
Consistent Synchronization Schemes for Workload Replay ..... <i>Konstantinos Morfonios, Romain Colle, Leonidas Galanis, Supiti Buranawanachoke, Benoît Dageville, Karl Dias, Yujun Wang</i>	1225
Inspector Gadget: A Framework for Custom Monitoring and Debugging of Distributed Dataflows ..... <i>Christopher Olston, Benjamin Reed</i>	1237
Online Expansion of Large-scale Data Warehouses ..... <i>Jeffrey Cohen, John Eshleman, Brian Hagenbuch, Joy Kent, Christopher Pedrotti, Gavin Sherry, Florian Waas</i>	1249
HIWAS: Enabling Technology for Analysis of Clinical Data in XML Documents ..... <i>Joshua Hui, Sarah Knoop, Peter Schwarz</i>	1260
Jaql: A Scripting Language for Large Scale Semi-Structured Data Analysis ..... <i>K. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M. Eltabakh, C.-C. Kanne, F. Ozcan, E. Shekita</i>	1272
Auto-Grouping Emails for Faster E-Discovery ..... <i>Sachindra Joshi, Danish Contractor, Kenney Ng, Prasad Deshpande, Thomas Hampf</i>	1284
Web Scale Taxonomy Cleansing ..... <i>Taesung Lee, Zhongyuan Wang, Haixun Wang, Seung-won Hwang</i>	1295
Bridging Two Worlds with RICE ..... <i>Philipp Große, Wolfgang Lehner, Thomas Weichert, Franz Färber, Wen-Syan Li</i>	1307
Tenzing – A SQL Implementation On The MapReduce Framework..... <i>B. Chattopadhyay, L. Lin, W. Liu, S. Mittal, P. Aragonda, V. Lychagina, Y. Kwon, M. Wong</i>	1318
An Algebraic Approach for Data-Centric Scientific Workflows ..... <i>Eduardo Ogasawara, Jonas Dias, Daniel de Oliveira, Fábio Porto, Patrick Valduriez, Marta Mattoso</i>	1328
Citrusleaf: A Real-Time NoSQL DB which Preserves ACID ..... <i>V. Srinivasan, Brian Bulkowski</i>	1340

## Demonstration Track Papers

RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows .....	1351
..... <i>Hyunjung Park, Robert Ikeda, Jennifer Widom</i>	
BROAD: Diversified Keyword Search in Databases .....	1355
..... <i>Feng Zhao, Xiaolong Zhan, Anthony Tung, Gang Chen</i>	
TrustedDB: A Trusted Hardware based Database with Privacy and Data Confidentiality .....	1359
..... <i>Sumeet Bajaj, Radu Sion</i>	
IPL-P: In-Page Logging with PCRAM .....	1363
..... <i>Kang-Nyeon Kim, Sang-Won Lee, Bongki Moon, Chanik Park, Joo-Young Hwang</i>	
HyPer-sonic Combined Transaction AND Query Processing .....	1367
..... <i>Florian Funke, Alfons Kemper, Thomas Neumann</i>	
GrouPeer: A System for Clustering PDMSs.....	1371
..... <i>Verena Kantere, Dimos Bousounis, Timos Sellis</i>	
CerFix: A System for Cleaning Data with Certain Fixes .....	1375
..... <i>Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu</i>	
Online Visualization of Geospatial Stream Data using the WorldWide Telescope .....	1379
<i>Mohamed Ali, Badrish Chandramouli, Jonathan Fay, Curtis Wong, Steven Drucker, Balan S. Raman</i>	
Debugging Data Exchange with Vagabond .....	1383
..... <i>Boris Glavic, Jiang Du, Renée J. Miller, Gustavo Alonso, Laura M. Haas</i>	
CrowdDB: Query Processing with the VLDB Crowd.....	1387
..... <i>Amber Feng, Michael Franklin, Donald Kossmann, Tim Kraska, Samuel Madden, Sukriti Ramesh, Andrew Wang, Reynold Xin</i>	
Analytics for the Real-Time Web.....	1391
..... <i>Maxim Grinev, Maria Grineva, Martin Hentschel, Donald Kossmann</i>	
DivDB: A System for Diversifying Query Results.....	1395
<i>.Maria C. Barioni, Marios Hadjieleftheriou, Divesh Srivastava, Caetano Traina Jr., Vassilis Tsotras</i>	
HOMES: A Higher-Order Mapping Evaluation System .....	1399
..... <i>Huy Vu, Michael Benedikt</i>	
Proactive Detection and Repair of Data Corruption: Towards a Hassle-free Declarative Approach with Amulet .....	1403
..... <i>Nedyalko Borisov, Shivnath Babu</i>	
Automatic Workload Driven Index Defragmentation .....	1407
..... <i>Vivek Narasayya, Hyunjung Park, Manoj Syamala</i>	
Whodunit: An Auditing Tool for Detecting Data Breaches .....	1410
..... <i>Raghav Kaushik, Ravi Ramamurthy</i>	

EIRENE: Interactive Design and Refinement of Schema Mappings via Data Examples.....	1414
..... <i>Bogdan Alexe, Balder ten Cate, Phokion G. Kolaitis, Wang-Chiew Tan</i>	
DataSynth: Generating Synthetic Data using Declarative Constraints.....	1418
..... <i>Arvind Arasu, Raghav Kaushik, Jian Li</i>	
InfoNetOLAPer: Integrating InfoNetWarehouse and InfoNetCube with InfoNetOLAP .....	1422
..... <i>Chuan Li, Philip S. Yu, Lei Zhao, Yan Xie, Wangqun Lin</i>	
From SPARQL to MapReduce: The Journey Using a Nested TripleGroup Algebra .....	1426
..... <i>HyeongSik Kim, Padmashree Ravindra, Kemafor Anyanwu</i>	
FuDoCS: A Web Service Composition System Based on Fuzzy Dominance for Preference Query	1430
Answering..... <i>Karim Benouaret, Djamel Benslimane, Allel Hadjali, Mahmoud Barhamgi</i>	
A Demonstration of HYRISE – A Main Memory Hybrid Storage Engine .....	1434
..... <i>Martin Grund, Philippe Cudre-Mauroux, Samuel Madden</i>	
++Spicy: an Open-Source Tool for Second-Generation Schema Mapping and Data Exchange .....	1438
..... <i>Bruno Marnette, Giansalvatore Mecca, Paolo Papotti, Salvatore Raunich, Donatello Santoro</i>	
UpStream: A Storage-centric Load Management System for Real-time Update Streams .....	1442
..... <i>Alexandru Moga, Nesime Tatbul</i>	
MapReduce Programming and Cost-based Optimization? Crossing this Chasm with Starfish .....	1446
..... <i>Herodotos Herodotou, Fei Dong, Shivnath Babu</i>	
AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables.....	1450
..... <i>Mohamed Amir Yosef, Johannes Hoffart, Iliaria Bordino, Marc Spaniol, Gerhard Weikum</i>	
Microsoft Codename “Montego” – Data Import, Transformation, and Publication for Information	1454
Workers..... <i>Stephen J. Maine, Lorenz Prem, Clemens Szyperski, James F. Terwilliger</i>	
SocialSpamGuard: A Data Mining-Based Spam Detection System for Social Media Networks .....	1458
..... <i>Xin Jin, Cindy Xide Lin, Jiebo Luo, Jiawei Han</i>	

### Challenges and Vision Track Papers

Resiliency-Aware Data Management .....	1462
..... <i>Matthias Böhm, Wolfgang Lehner, Christof Fetzer</i>	
Guided Interaction: Rethinking the Query-Result Paradigm.....	1466
..... <i>Arnab Nandi, H. V. Jagadish</i>	
Data Generation for Application-Specific Benchmarking .....	1470
..... <i>Y. C. Tay</i>	
The Researcher's Guide to the Data Deluge: Querying a Scientific Database in Just a Few Seconds	1474
..... <i>Martin L. Kersten, Stratos Idreos, Stefan Manegold, Erietta Liarou</i>	

Anthropocentric Data Systems.....	1478
..... <i>Peter Triantafillou</i>	
Data Markets in the Cloud: An Opportunity for the Database Community.....	1482
..... <i>Magdalena Balazinska, Bill Howe, Dan Suciu</i>	
Data is Dead... Without What-if Models.....	1486
..... <i>Peter Haas, Paul P. Maglio, Patricia G. Selinger, Wang-Chiew Tan</i>	
Reverse Data Management .....	1490
..... <i>Alexandra Meliou, Wolfgang Gatterbauer, Dan Suciu</i>	
Exploring the Coming Repositories of Reproducible Experiments: Challenges and Opportunities ..	1494
..... <i>Juliana Freire, Philippe Bonnet, Dennis Shasha</i>	
Databases will Visualize Queries too .....	1498
..... <i>Wolfgang Gatterbauer</i>	

## Tutorials

New Frontiers in Business Intelligence .....	1502
..... <i>Surajit Chaudhuri, Vivek Narasayya</i>	
System Co-Design and Data Management for Flash Devices .....	1504
..... <i>Philippe Bonnet, Luc Bouganim, Ioannis Koltsidas, Stratis D. Viglas</i>	
Exploration of Deep Web Repositories.....	1506
..... <i>Nan Zhang, Gautam Das</i>	
Crowdsourcing Applications and Platforms: A Data Management Perspective .....	1508
..... <i>AnHai Doan, Michael J. Franklin, Donald Kossmann, Tim Kraska</i>	
Graph Data Management Systems for New Application Domains .....	1510
..... <i>Philippe Cudré-Mauroux, Sameh Elnikety</i>	
Information Diffusion in Social Networks: Observing and Influencing Societal Interests .....	1512
..... <i>Divyakant Agrawal, Ceren Budak, Amr El Abbadi</i>	

## Panels

Data Management for Meeting Global Health Challenges .....	1514
..... <i>Tapan S. Parikh, Kuang Chen</i>	
Panel Discussion: Maximizing Impact .....	1515
..... <i>Ed Lazowska</i>	

## VLDB 2011 PROGRAM COMMITTEES

### VLDB 2011 General PC Co-Chairs

José Blakeley, Microsoft

Joseph Hellerstein, University of California – Berkeley

### Industrial and Applications Track Co-Chairs

Berthold Reinwald, IBM Almaden Research Center

Phil Bohannon, Yahoo! Research

### Industrial and Applications Track Program Committee

Roger Barga (Microsoft Research)

Christof Bornhoevd (SAP Palo Alto)

Brian Cooper (Google)

Anish Das Sarma (Yahoo! Research)

Mario Inghiosa (Netezza)

Avinash Lakshman (Facebook)

Zhen Hua Liu (Oracle)

Jun Rao (LinkedIn)

Mehul Shah (HP Labs)

Adam Silberstein (Yahoo! Research)

Alkis Simitsis (HP Labs)

Garret Swart (Oracle)

Yuanyuan Tian (IBM Research)

Yu Xu (Teradata)

Peter Zabback (Microsoft)

Calisto Zuzarte (IBM)

### Demonstration Track Co-Chairs

Jignesh Patel, University of Wisconsin – Madison

Masatoshi Yoshikawa, University of Kyoto

### Demonstration Track Program Committee

Sourav Bhowmick (Nanyang Technological University, Singapore)

Dhruba Borthakur (Facebook)

Sarah Boulakia (LRI Orsay, France)

Sang Cha (Seoul National University, Korea)

Venkatesh Ganti (Google)

Shahram Ghandeharizadeh (U. Southern California)

Yoshiharu Ishikawa (Nagoya University, Japan)

Volker Markl (TU Berlin, Germany)

Jun Miyazaki (Nara Institute of Science and Technology)

Bongki Moon (University of Arizona)

Atsuyuki Morishima (Tsukuba University, Japan)

Rimma Nehme (Microsoft Research)

Jun Rao (LinkedIn)

Rajeev Rastogi (Yahoo! Research India)

Keishi Tajima (Kyoto University, Japan)

Xiaofang Zhou (University of Queensland, Australia)

**Challenges and Vision Track Chair**

Gerhard Weikum, Max Planck Institute for Informatics

**Challenges and Vision Track Program Committee**

Anastassia Ailamaki (EPF Lausanne)

Sihem Amer-Yahia (Yahoo! Research)

Philip Bernstein (Microsoft Research)

Michael Cafarella (University of Michigan)

Susan Davidson (University of Pennsylvania)

Laura Haas (IBM Almaden Research Center)

Meichun Hsu (HP Labs)

Hank Korth (Lehigh University)

**VLDB 2011 Tutorial Program Co-Chairs**

Qiong Luo, Hong Kong University of Science & Techn.

Gerome Miklau, University of Massachusetts – Amherst

**PVLDB Information Director**

Gerald Weber (University of Auckland)

**VLDB 2011 Proceedings Chair**

Uwe Röhm (University of Sydney)

**Steering Committee**

Serge Abiteboul, Peter Apers, Philip Bernstein, Elisa Bertino, Peter Buneman, Martin Kersten, Z. Meral Ozsoyuglu



## LETTER FROM THE INDUSTRIAL TRACK CHAIRS

As data volumes increase dramatically, and grid- and cloud-processing techniques are adopted by more and more industrial and government entities, the challenges for processing, migrating and searching this data are multiplying. As a result, a variety of platforms, development environments and management tools are required, each with a unique set of technical challenges. This year's Industrial Track of VLDB is a snapshot of this development from several areas of industry. In particular, these proceedings include several papers on managing workflow, programming environments, debugging and analysis task on grid frameworks. In addition, there is interesting work on replaying workloads for testing, data warehousing, clinical data management, email processing and data cleaning. The papers were drawn from a strong body of submissions, and on behalf of the program committee, we would like to thank the authors for the obvious effort put into their submissions.

---

Berthold Reinwald, IBM Almaden Research Center  
Phil Bohannon, Yahoo! Research  
VLDB 2011 Industrial, Applications, and Experience Track Chairs