

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Alfred Kobsa

*University of California, Irvine, CA, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Germany*

Madhu Sudan

*Microsoft Research, Cambridge, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max-Planck Institute of Computer Science, Saarbruecken, Germany*

Michael Gertz Bertram Ludäscher (Eds.)

# Scientific and Statistical Database Management

22nd International Conference, SSDBM 2010  
Heidelberg, Germany, June 30–July 2, 2010  
Proceedings

Volume Editors

Michael Gertz  
University of Heidelberg  
Institute of Computer Science  
69120 Heidelberg, Germany  
E-mail: gertz@informatik.uni-heidelberg.de

Bertram Ludäscher  
University of California  
Dept. of Computer Science and Genome Center  
One Shields Avenue, Davis, CA 95616, USA  
E-mail: ludaesch@ucdavis.edu

Library of Congress Control Number: 2010929609

CR Subject Classification (1998): H.3, I.2, H.4, C.2, H.2.8, H.5

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743  
ISBN-10 3-642-13817-9 Springer Berlin Heidelberg New York  
ISBN-13 978-3-642-13817-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2010  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper 06/3180

# Preface

The International Conference on Scientific and Statistical Database Management (SSDBM) is an established forum for the exchange of the latest research results on concepts, tools, and techniques for scientific database applications. The 2010 meeting marked the 22<sup>nd</sup> time that scientific domain experts, databases researchers, practitioners, and developers came together to share their insights and to discuss future research directions in a stimulating environment. The conference was held from June 30 to July 2 at Villa Bosch, near the Carl Bosch Museum and Heidelberg Castle, overlooking the picturesque Neckar Valley. The conference was organized at and co-sponsored by Heidelberg University and HITS, the Heidelberg Institute for Theoretical Studies, established in January 2010 by Dr. Klaus Tschira, co-founder of SAP AG, as a successor to the EML Research Institute. HITS focuses on new approaches and foundations towards interpreting the rapidly increasing amounts of experimental data. Heidelberg University, the oldest university in Germany, was founded in 1386, is a German Excellence University, and a top-ranking university in Europe and worldwide, known among other things for its excellence in the natural sciences and medicine. The university also hosts a unique Interdisciplinary Center for Scientific Computing (IWR) where grand challenges in the computational sciences are tackled, e.g., climate and ocean modeling, turbulent flows, combustion, bio-molecules, and drug design.

In 2010, SSDBM received a near-record number of 94 submissions from 27 countries. Each submission was reviewed by at least three of the 38 PC members or external reviewers. After careful consideration, 41 papers ( $\approx 44\%$ ) overall were accepted, 30 as long papers ( $\approx 32\%$ ) and 11 short papers and demonstrations. The reviewing process was managed by the EasyChair Conference System ([www.easychair.org](http://www.easychair.org)), an excellent free conference management system, developed by Andrei Voronkov. SSDBM 2010 featured two keynotes: Daniel Abadi from Yale University discussed “Trade-offs Between Parallel Database Systems, Hadoop, and HadoopDB as Platforms for Petabyte-Scale Analysis” and described experiences using existing parallel databases and MapReduce systems, and HadoopDB, a hybrid system under development at Yale. In his keynote, Roger Barga from Microsoft Research presented “Emerging Trends and Converging Technologies in Data-Intensive Scalable Computing.”

The program and activities of SSDBM 2010 were the result of a large effort by the authors, reviewers, presenters, and organizers. We thank them all for helping to make this conference a success! In particular, we thank the General Chair, Andreas Reuter for offering to host SSDBM at Villa Bosch, for supporting SSDBM in general, and for organizing the SSDBM sponsors. We also thank Wolfgang Müller from HITS for the local organization and Conny Franke for compiling the proceedings. Last not least, we would like to thank the “father of SSDBM”

and Chair of the Steering Committee, Arie Shoshani, for guidance throughout the process of organizing SSDBM. We hope you enjoy the proceedings!

April 2010

Michael Gertz  
Bertram Ludäscher

# Conference Organization

## General and Program Chairs

Andreas Reuter (General Chair)  
Michael Gertz (PC Chair)  
Tony Hey (PC Co-chair)  
Bertram Ludäscher (PC Co-chair)

## Local Organization

Wolfgang Müller

## Proceedings Chair

Conny Franke

## Program Committee

Ilkay Altintas  
Walid G. Aref  
Elisa Bertino  
Shawn Bowers  
Isabel F. Cruz  
Conny Franke  
Juliana Freire  
Johann Gamper  
Floris Geerts  
Amarnath Gupta  
Ralf Hartmut Güting  
Bill Howe  
Theo Härder  
Christian S. Jensen  
Martin Kersten  
Peer Kröger  
Zoé Lacroix  
Ulf Leser  
David Maier  
Paolo Missier  
Mohamed F. Mokbel  
Suman Nath  
Silvia Nittel

Dimitris Papadias  
Panagiotis Papapetrou  
Norman W. Paton  
Tore Risch  
Carlos Rueda  
Nagiza F. Samatova  
Thomas Seidl  
Rishi Rakesh Sinha  
Kurt Stockinger  
Alex S. Szalay  
Yufei Tao  
Nesime Tatbul  
Can Türker  
Jianwu Wang  
Daniel Zinn

## External Reviewers

Ira Assent  
Sebastian Bächle  
Jie Bao  
Roger Barga  
Christian Beecks  
Thomas Behr  
Khalid Belhajjame  
Thomas Bernecker  
Roger Castillo  
Vasa Curcin  
Biplob Debnath  
Robin Dhamankar  
Tobias Emrich  
Ines Färber  
Alvaro A.A. Fernandes  
Sergej Fries  
Ixent Galpin  
Gabriel Ghinita  
Rigel Gjomemo  
Keith Grochow  
Alasdair J.G. Gray  
Todd Green  
Stephan Günemann  
Cornelia Hedeler  
Abdeltawab Hendawi  
Volker Hudlet  
Hoyoung Jeung

Mouna Kacimi  
Murat Kantarcioglu  
Mohamed Khalefa  
Kraig King  
Andre Koschmieder  
Hardy Kremer  
YongChul Kwon  
Xiang Lian  
Qinghan Liang  
Jefrey Lijffijt  
Hyo-Sang Lim  
Hua Lu  
Angela Maduko  
Soumyadeb Mitra  
Emmanuel Müller  
Mohamed Nabeel  
Kjell Orsborn  
Stavros Papadopoulos  
Astrid Rheinländer  
Dimitris Sacharidis  
Satya Sahoo  
Mahmoud Sakr  
Daniel Schall  
Karsten Schmidt  
Christian Sengstock  
Silvia Stefanova  
Arash Termehchy

George Trimponias  
Silke Trissl  
Lixing Wang  
Andreas Weiner  
Marc Wichterich  
Dingming Wu

Yin Yang  
Ying Yang  
Jun Zhao  
Erik Zeitler  
Andreas Züfle



# Table of Contents

## Invited Talks

Tradeoffs between Parallel Database Systems, Hadoop, and HadoopDB as Platforms for Petabyte-Scale Analysis . . . . .	1
<i>Daniel J. Abadi</i>	

Emerging Trends and Converging Technologies in Data Intensive Scalable Computing . . . . .	4
<i>Roger S. Barga</i>	

## Query Processing

Deriving Spatio-temporal Query Results in Sensor Networks . . . . .	6
<i>Markus Besthorn, Klemens Böhm, Patrick Bradley, and Erik Buchmann</i>	

Efficient and Adaptive Distributed Skyline Computation . . . . .	24
<i>George Valkanas and Apostolos N. Papadopoulos</i>	

On the Efficient Construction of Multislices from Recurrences . . . . .	42
<i>Romans Kasperovics, Michael H. Böhlen, and Johann Gamper</i>	

Optimizing Query Processing in Cache-Aware Wireless Sensor Networks . . . . .	60
<i>Mario A. Nascimento, Romulo A.E. Alencar, and Angelo Brayner</i>	

Approximate Query Answering and Result Refinement on XML Data . . .	78
<i>Katja Seidler, Eric Peukert, Gregor Hackenbroich, and Wolfgang Lehner</i>	

Efficient and Scalable Method for Processing Top-k Spatial Boolean Queries . . . . .	87
<i>Ariel Cary, Ouri Wolfson, and Naphtali Rish</i>	

## Scientific Data Management and Analysis

A Framework for Moving Sensor Data Query and Retrieval of Dynamic Atmospheric Events . . . . .	96
<i>Shen-Shyang Ho, Wenqing Tang, W. Timothy Liu, and Markus Schneider</i>	

Client + Cloud: Evaluating Seamless Architectures for Visual Data Analytics in the Ocean Sciences . . . . .	114
<i>Keith Grochow, Bill Howe, Mark Stoermer, Roger Barga, and Ed Lazowska</i>	
Scalable Clustering Algorithm for N-Body Simulations in a Shared-Nothing Cluster . . . . .	132
<i>YongChul Kwon, Dylan Nunley, Jeffrey P. Gardner, Magdalena Balazinska, Bill Howe, and Sarah Loebman</i>	
Database Design for High-Resolution LIDAR Topography Data . . . . .	151
<i>Viswanath Nandigam, Chaitan Baru, and Christopher Crosby</i>	
PetaScope: An Open-Source Implementation of the OGC WCS Geo Service Standards Suite . . . . .	160
<i>Andrei Aiordăchioaie and Peter Baumann</i>	
Towards Archaeo-Informatics: Scientific Data Management for Archaeobiology . . . . .	169
<i>Hans-Peter Kriegel, Peer Kröger, Christiaan Hendrikus van der Meijden, Henriette Obermaier, Joris Peters, and Matthias Renz</i>	
<b>Data Mining</b>	
DESSIN: Mining Dense Subgraph Patterns in a Single Graph . . . . .	178
<i>Shirong Li, Shijie Zhang, and Jiong Yang</i>	
Discovery of Evolving Convoys . . . . .	196
<i>Htoo Htet Aung and Kian-Lee Tan</i>	
Finding Top-k Similar Pairs of Objects Annotated with Terms from an Ontology . . . . .	214
<i>Arnab Bhattacharya, Abhishek Bhowmick, and Ambuj K. Singh</i>	
Identifying the Most Influential User Preference from an Assorted Collection . . . . .	233
<i>Hua Lu and Linhao Xu</i>	
MC-Tree: Improving Bayesian Anytime Classification . . . . .	252
<i>Philipp Kranen, Stephan Günemann, Sergej Fries, and Thomas Seidl</i>	
Non-intrusive Quality Analysis of Monitoring Data . . . . .	270
<i>Mark Brightwell, Anastasia Ailamaki, and Anna Suwalska</i>	

Visual Decision Support for Ensemble Clustering . . . . .	279
<i>Martin Hahmann, Dirk Habich, and Wolfgang Lehner</i>	

## Indexes and Data Representation

An Indexing Scheme for Fast and Accurate Chemical Fingerprint Database Searching . . . . .	288
<i>Zeyar Aung and See-Kiong Ng</i>	
BEMC: A Searchable, Compressed Representation for Large Seismic Wavefields . . . . .	306
<i>Julio López, Leonardo Ramírez-Guzmán, Jacobo Bielak, and David O'Hallaron</i>	
Dynamic Data Reorganization for Energy Savings in Disk Storage Systems . . . . .	322
<i>Ekow Otoo, Doron Rotem, and Shih-Chiang Tsao</i>	
Organization of Data in Non-convex Spatial Domains . . . . .	342
<i>Eric Perlman, Randal Burns, Michael Kazhdan, Rebecca R. Murphy, William P. Ball, and Nina Amenta</i>	
PrefIndex: An Efficient Supergraph Containment Search Technique . . . . .	360
<i>Gaoping Zhu, Xuemin Lin, Wenjie Zhang, Wei Wang, and Haichuan Shang</i>	
Supporting Web-Based Visual Exploration of Large-Scale Raster Geospatial Data Using Binned Min-Max Quadtree . . . . .	379
<i>Jianting Zhang and Simin You</i>	

## Scientific Workflow and Provenance

Bridging Workflow and Data Provenance Using Strong Links . . . . .	397
<i>David Koop, Emanuele Santos, Bela Bauer, Matthias Troyer, Juliana Freire, and Cláudio T. Silva</i>	
LIVE: A Lineage-Supported Versioned DBMS . . . . .	416
<i>Anish Das Sarma, Martin Theobald, and Jennifer Widom</i>	
Optimizing Resource Allocation for Scientific Workflows Using Advance Reservations . . . . .	434
<i>Christoph Langguth and Heiko Schuldts</i>	
A Fault-Tolerance Architecture for Kepler-Based Distributed Scientific Workflows . . . . .	452
<i>Pierre Moullem, Daniel Crawl, Ilkay Altintas, Mladen Vouk, and Ustun Yildiz</i>	

Provenance Context Entity (PaCE): Scalable Provenance Tracking for Scientific RDF Data . . . . . 461  
*Satya S. Sahoo, Olivier Bodenreider, Pascal Hitzler, Amit Sheth, and Krishnaprasad Thirunarayan*

Taverna, Reloaded . . . . . 471  
*Paolo Missier, Stian Soiland-Reyes, Stuart Owen, Wei Tan, Alexandra Nenadic, Ian Dunlop, Alan Williams, Tom Oinn, and Carole Goble*

**Similarity**

Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? . . . 482  
*Michael E. Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek*

Optimizing All-Nearest-Neighbor Queries with Trigonometric Pruning . . . . . 501  
*Tobias Emrich, Franz Graf, Hans-Peter Kriegel, Matthias Schubert, and Marisa Thoma*

Prefix Tree Indexing for Similarity Search and Similarity Joins on Genomic Data . . . . . 519  
*Astrid Rheinländer, Martin Knobloch, Nicky Hochmuth, and Ulf Leser*

Similarity Estimation Using Bayes Ensembles . . . . . 537  
*Tobias Emrich, Franz Graf, Hans-Peter Kriegel, Matthias Schubert, and Marisa Thoma*

Subspace Similarity Search: Efficient k-NN Queries in Arbitrary Subspaces . . . . . 555  
*Thomas Bernecker, Tobias Emrich, Franz Graf, Hans-Peter Kriegel, Peer Kröger, Matthias Renz, Erich Schubert, and Arthur Zimek*

**Data Stream Processing**

Continuous Skyline Monitoring over Distributed Data Streams . . . . . 565  
*Hua Lu, Yongluan Zhou, and Jonas Haustad*

Propagation of Densities of Streaming Data within Query Graphs . . . . . 584  
*Michael Daum, Frank Lauterwald, Philipp Baumgärtel, and Klaus Meyer-Wegener*

Spatio-temporal Event Stream Processing in Multimedia Communication Systems . . . . . 602  
*Mingyan Gao, Xiaoyan Yang, Ramesh Jain, and Beng Chin Ooi*

Stratified Reservoir Sampling over Heterogeneous Data Streams . . . . .	621
<i>Mohammed Al-Kateb and Byung Suk Lee</i>	
Tree Induction over Perennial Objects . . . . .	640
<i>Zaigham Faraz Siddiqui and Myra Spiliopoulou</i>	
<b>Author Index</b> . . . . .	659