Preface

Data warehousing and knowledge discovery technologies have been growing over time and they are now a practical need in every major small and large company. The maturity of these technologies encourages these companies to integrate heterogeneous, autonomous and evolving data issued from traditional (databases) and advanced sources such as sensors and social networks into a single large database to enable advanced querying, analysis, and recommendation. The data warehouse design passes through a well-established lifecycle that includes: conceptual modeling, ETL, logical modeling, deployment, and physical modeling. Knowledge discovery complements this lifecycle to offering exploitation capabilities of the warehouse. With the diversity of deployment platforms motivated by HPC (High Processing Computing), offering the process of retrieval and knowledge discovery from this huge amount of heterogeneous complex data builds the litmus-test for the research in the area.

During the past years, the International Conference on Data Warehousing and Knowledge Discovery (DaWaK) has become one of the most important international scientific events to bring together researchers, developers, and practitioners to discuss latest research issues and experiences in developing and deploying data warehousing and knowledge discovery systems, applications, and solutions. DaWaK is in the top 20 of the google scholar ranking related to Data Mining & Analysis: http://scholar.google.com/citations?view_op=top_venues&hl=fr&vq=eng_datamininganalysis. This year’s conference (DaWaK 2014), builds on this tradition of facilitating the cross-disciplinary exchange of ideas, experience, and potential research directions. DaWaK 2014 seeks to introduce innovative principles, methods, models, algorithms and solutions, industrial products, and experiences to challenging problems faced in the development of data warehousing, knowledge discovery, data mining applications, and the emerging area of HPC.

This year we received 109 papers and the Program Committee finally selected 34 full papers and 8 short papers, making an acceptance rate of 31%. The accepted papers cover a number of broad research areas on both theoretical and practical aspects of data warehouse and knowledge discovery. In the area of data warehousing, the topics covered included the modeling and ETL, ontologies, real-time data warehouses, query optimization, map reduce paradigm, storage models, scalability, distributed and parallel processing and data warehouses and data mining applications integration, recommendation and personalization, multidimensional analysis of text documents, and data warehousing for real world applications such as health, bio-informatics, telecommunication, etc. In the areas of data mining and knowledge discovery, the topics included stream data analysis and mining, traditional data mining techniques topics such as frequent item sets, clustering, association, classification ranking and application of data mining technologies to real world problems, and fuzzy mining, skyline, etc.
It is especially notable to see that some papers covered emerging real world applications as bioinformatics, social network, telecommunication, brain analysis, etc.


We would like to thank all authors for submitting their research paper in DaWaK 2014. We express our gratitude to all the Program Committee members and the external reviews, who reviewed the papers very profoundly and in a timely manner. Finally, we would like to thank Mrs Gabriela Wagner for her endless help and support.

See you all in Munich. Hope you enjoy the technical program, meeting and interacting with research colleagues, and of course, a beautiful city Munich.

July 2014

Ladjel Bellatreche
Mukesh K. Mohania
Organization

Program Committee Co-chairs

Ladjel Bellatreche  LIAS/ISAE-ENSMA, France
Mukesh Mohania    IBM Research, India

Program Committee

Alberto Abelló    Universitat Politecnica de Catalunya, Spain
Rajeev Agrawal   North Carolina A&T State University, USA
Mohammed Al-Kateb Teradata Labs, USA
Torben Bach Pedersen Aalborg University, Denmark
Elena Baralis    Politecnico di Torino, Italy
Sadok Ben Yahia  Faculty of Sciences of Tunis, France
Petr Berka       University of Economics in Prague, Czech Republic
Jorge Bernardino ISEC - Polytechnic Institute of Coimbra, Portugal
Vasudha Bhatnagar Delhi University, India
Manish Bhide     IBM Software Lab, India
Kamel Boukhalfa  USTHB, Algeria
Omar Boussaid    University of Lyon, France
Stephane Bressan National University of Singapore, Singapore
Erik Buchmann    Karlsruhe Institute of Technology, Germany
Frans Coenen     The University of Liverpool, UK
Bruno Cremilleux Université de Caen, France
Alfredo Cuzzocrea ICAR-CNR and University of Calabria, Italy
Karen Davis      University of Cincinnati, USA
Alin Dobra       University of Florida, USA
Dejing Dou       University of Oregon, USA
Curtis Dyreson   Utah State University, USA
Markus Endres    University of Augsburg, Germany
Vladimir Estivill-Castro Griffith University, Australia
Christie Ezeife  University of Windsor, Canada
Filippo Furfaro   University of Calabria, Italy
Pedro Furtado    Universidade de Coimbra, Portugal
Carlos Garcia-Alvarado Pivotal Inc., USA
Matteo Golfarelli DISI - University of Bologna, Italy
Sergio Greco     University of Calabria, Italy
Neelima Gupta    University of Delhi, India
VIII Organization

Frank Hoppner
Ostfalia University of Applied Sciences,
Germany

Domingo-Ferrer Josep
University Rovira i Virgili, Spain

Selma Khouri
National High School of Computer Science,
Algeria

Sang-Wook Kim
Hanyang University, South Korea

Christian Koncilia
University of Klagenfurt, Austria

Jens Lechtenboerger
Westfälische Wilhelms-Universität Münster,
Germany

Wolfgang Lehner
Dresden University of Technology, Germany

Carson K. Leung
The University of Manitoba, Canada

Sofian Maabout
University of Bordeaux, France

Patrick Marcel
Université François Rabelais Tours, France

Anirban Mondal
Xerox Research, India

Apostolos Papadopoulos
Aristotle University, Greece

Jeffrey Parsons
Memorial University of Newfoundland, Canada

Dhaval Patel
Indian Institute of Technology Roorkee, India

Lu Qin
University of Technology in Sydney, Australia

Sriram Raghavan
IBM Research - India

Stefano Rizzi
University of Bologna, Italy

Maria Luisa Sapino
Università degli Studi di Torino, Italy

Kai-Uwe Sattler
Ilmenau University of Technology, Germany

Neeraj Sharma
India IBM Labs, India

Alkis Simitsis
HP Labs, USA

Domenico Talia
University of Calabria, Italy

David Taniar
Monash University, Australia

Olivier Teste
IRIT, University of Toulouse, France

Dimitri Theodoratos
New Jersey Institute of Technology, USA

A Min Tjoa
Vienna University of Technology, Austria

Panos Vassiliadis
University of Ioannina, Greece

Robert Wrembel
Poznan University of Technology, Poland

Wolfram Wöß
Johannes Kepler University Linz, Austria

Karine Zeitouni
Université de Versailles
Saint-Quentin-en-Yvelines, France

Bin Zhou
University of Maryland in Baltimore County, USA

External Reviewers

Christian Thomsen
Aalborg University, Denmark

Xike Xie
Aalborg University, Denmark

Anamika Gupta
CBS College, University of Delhi, India

Rakhi Saxena
Deshbandhu College, University of Delhi, India

Florian Wenzel
University of Augsburg, Germany

Patrick Roocks
University of Augsburg, Germany
A Secure Data Sharing and Query Processing Framework via Federation of Cloud Computing (Keynote)

Sanjay K. Madria

Department of Computer Science
Missouri University of Science and Technology, Rolla, MO, USA
madrias@mst.edu

Abstract. Due to cost-efficiency and less hands-on management, big data owners are outsourcing their data to the cloud, which can provide access to the data as a service. However, by outsourcing their data to the cloud, the data owners lose control over their data, as the cloud provider becomes a third party service provider. At first, encrypting the data by the owner and then exporting it to the cloud seems to be a good approach. However, there is a potential efficiency problem with the outsourced encrypted data when the data owner revokes some of the users’ access privileges. An existing solution to this problem is based on symmetric key encryption scheme but it is not secure when a revoked user rejoins the system with different access privileges to the same data record. In this talk, I will discuss an efficient and Secure Data Sharing (SDS) framework using a combination of homomorphic encryption and proxy re-encryption schemes that prevents the leakage of unauthorized data when a revoked user rejoins the system. I will also discuss the modifications to our underlying SDS framework and present a new solution based on the data distribution technique to prevent the information leakage in the case of collusion between a revoked user and the cloud service provider. A comparison of the proposed solution with existing methods will be discussed. Furthermore, I will outline how the existing work can be utilized in our proposed framework to support secure query processing for big data analytics. I will provide a detailed security as well as experimental analysis of the proposed framework on Amazon EC2 and highlight its practical use.

Biography: Sanjay Kumar Madria received his Ph.D. in Computer Science from Indian Institute of Technology, Delhi, India in 1995. He is a full professor in the Department of Computer Science at the Missouri University of Science and Technology (formerly, University of Missouri-Rolla, USA) and site director, NSF I/UCRC center on Net-Centric Software Systems. He has published over 200 Journal and conference papers in the areas of mobile data management, Sensor computing, and cyber security and trust management. He won three best papers awards including IEEE MDM 2011 and IEEE MDM 2012. He is the co-author of a book published by Springer in Nov 2003. He serves as steering committee members in IEEE SRDS and IEEE MDM among others and has served in International conferences as a general co-chair (IEEE MDM,
IEEE SRDS and others), and presented tutorials/talks in the areas of mobile data management and sensor computing at various venues. His research is supported by several grants from federal sources such as NSF, DOE, AFRL, ARL, ARO, NIST and industries like Boeing, Unique*Soft, etc. He has also been awarded JSPS (Japanese Society for Promotion of Science) visiting scientist fellowship in 2006 and ASEE (American Society of Engineering Education) fellowship at AFRL from 2008 to 2012. In 2012-13, he was awarded NRC Fellowship by National Academies. He has received faculty excellence research awards in 2007, 2009, 2011 and 2013 from his university for excellence in research. He served as an IEEE Distinguished Speaker, and currently, he is an ACM Distinguished Speaker, and IEEE Senior Member and Golden Core awardee.
Table of Contents

Modeling and ETL

An Approach on ETL Attached Data Quality Management ............... 1
Christian Lettner, Reinhard Stumptner, and Karl-Heinz Bokesch

Quality Measures for ETL Processes ........................................ 9
Vasileios Theodorou, Alberto Abelló, and Wolfgang Lehner

A Logical Model for Multiversion Data Warehouses ...................... 23
Waqas Ahmed, Esteban Zimányi, and Robert Wrembel

Ontology-Based Data Warehouses

OntoWarehousing – Multidimensional Design Supported by a Foundational Ontology: A Temporal Perspective .................. 35
João Moreira, Kelli Cordeiro, Maria Luiza Campos, and Marcos Borges

Modeling and Querying Data Warehouses on the Semantic Web Using QB4OLAP ................................................................. 45
Lorena Etcheverry, Alejandro Vaisman, and Esteban Zimányi

Extending Drill-Down through Semantic Reasoning on Indicator Formulas................................................................. 57
Claudia Diamantini, Domenico Potena, and Emanuele Storti

Advanced Data Warehouses and OLAP

An OLAP-Based Approach to Modeling and Querying Granular Temporal Trends................................................................. 69
Alberto Sabaini, Esteban Zimányi, and Carlo Combi

Real-Time Data Warehousing: A Rewrite/Merge Approach ............ 78
Alfredo Cuzzocrea, Nickerson Ferreira, and Pedro Furtado

Towards Next Generation BI Systems: The Analytical Metadata Challenge ................................................................. 89
Jovan Varga, Oscar Romero, Torben Bach Pedersen, and Christian Thomsen
Uncertainty

Mining Fuzzy Contextual Preferences .............................................. 102
Sandra de Amo and Juliete A. Ramos Costa

BLIMP: A Compact Tree Structure for Uncertain Frequent Pattern Mining .......................................................... 115
Carson Kai-Sang Leung and Richard Kyle MacKinnon

Discovering Statistically Significant Co-location Rules in Datasets with Extended Spatial Objects ............................................ 124
Jundong Li, Osmar R. Zaïane, and Alvaro Osornio-Vargas

Preferences and Recommendation

Discovering Community Preference Influence Network by Social Network Opinion Posts Mining .............................. 136
Tamanna S. Mumu and Christie I. Ezeife

Computing Hierarchical Skyline Queries “On-the-Fly” in a Data Warehouse .......................................................... 146
Tassadit Bouadi, Marie-Odile Cordier, and René Quiniou

Impact of Content Novelty on the Accuracy of a Group Recommender System ......................................................... 159
Ludovico Boratto and Salvatore Carta

Query Performance and HPC

Optimizing Queue-Based Semi-Stream Joins with Indexed Master Data .......................................................... 171
M. Asif Naeem, Gerald Weber, Christof Lutteroth, and Gillian Dobbie

Parallel Co-clustering with Augmented Matrices Algorithm with Map-Reduce .......................................................... 183
Meng-Lun Wu and Chia-Hui Chang

Processing OLAP Queries over an Encrypted Data Warehouse Stored in the Cloud ......................................................... 195
Claudivan Cruz Lopes, Valéria Cesário Times, Stan Matwin, Ricardo Rodrigues Ciferri, and Cristina Dutra de Aguiar Ciferri

Cube and OLAP

Reducing Multidimensional Data .......................................................... 208
Faten Atigui, Franck Ravat, Jiefu Song, and Gilles Zurfluh
Towards an OLAP Environment for Column-Oriented Data
Khaled Dehdouh, Fadila Bentayeb, Omar Boussaid, and Nadia Kabachi

Interval OLAP: Analyzing Interval Data
Christian Koncilia, Tadeusz Morzy, Robert Wrembel, and Johann Eder

Optimization
Improving the Processing of DW Star-Queries under Concurrent Query Workloads
João Pedro Costa and Pedro Furtado

Building Smart Cubes for Reliable and Faster Access to Data
Daniel K. Antwi and Herna L. Viktor

Optimizing Reaction and Processing Times in Automotive Industry’s Quality Management: A Data Mining Approach
Thomas Leitner, Christina Feilmayr, and Wolfram Wöß

Classification
Using Closed n-set Patterns for Spatio-Temporal Classification
Saulius Samulevičius, Yoann Pitarch, and Torben Bach Pedersen

Short Text Classification Using Semantic Random Forest
Ameni Bouaziz, Christel Dartigues-Pallez, Célia da Costa Pereira, Frédéric Precioso, and Patrick Lloret

3-D MRI Brain Scan Classification Using a Point Series Based Representation
Akadej Udomchaiporn, Frans Coenen, Marta García-Fiñana, and Vanessa Sluming

Social Networks and Recommendation Systems
Mining Interesting “Following” Patterns from Social Networks
Fan Jiang and Carson Kai-Sang Leung

A Novel Approach Using Context-Based Measure for Matching Large Scale Ontologies
Warith Eddine Djeddi and Mohamed Tarek Khadir

ActMiner: Discovering Location-Specific Activities from Community-Authored Reviews
Sahisnu Mazumder, Dhaval Patel, and Sameep Mehta
# Knowledge Data Discovery

A Scalable Algorithm for Banded Pattern Mining in Multi-dimensional Zero-One Data .................................................. 345  
*Fatimah B. Abdullahi, Frans Coenen, and Russell Martin*

Approximation of Frequent Itemset Border by Computing Approximate Minimal Hypergraph Transversals .......................... 357  
*Nicolas Durand and Mohamed Quafafou*

Clustering Based on Sequential Multi-Objective Games ............ 369  
*Imen Heloulou, Mohammed Said Radjef, and Mohand Tahar Kechadi*

# Industrial Applications

Opening up Data Analysis for Medical Health Services: Cancer Survival Analysis with CARESS ................................. 382  
*David Korfkamp, Stefan Gudenkauf, Martin Rohde, Eunice Sirri, Joachim Kieschke, and H.-Jürgen Appelrath*

“Real-time” Instance Selection for Biomedical Data Classification ...... 394  
*Chongsheng Zhang, Roberto D’Ambrosio, and Paolo Soda*

Mining Churning Factors in Indian Telecommunication Sector Using Social Media Analytics ........................................... 405  
*Nitish Varshney and S.K. Gupta*

# Mining and Processing Data Stream

Drift Detector for Memory-Constrained Environments ................. 414  
*Timothy D. Robinson, David Tse Jung Huang, Yun Sing Koh, and Gillian Dobbie*

Mean Shift Clustering Algorithm for Data with Missing Values ...... 426  
*Loai AbdAllah and Ilan Shimshoni*

Mining Recurrent Concepts in Data Streams Using the Discrete Fourier Transform .................................................... 439  
*Sakthithasan Sripirakas and Russel Pears*

# Mining and Similarity

Semi-Supervised Learning to Support the Exploration of Association Rules ................................................................. 452  
*Veronica Oliveira de Carvalho, Renan de Padua, and Solange Oliveira Rezende*
Parameter-Free Extended Edit Distance ............................................. 465
  Muhammad Marwan Muhammad Fuad

VGEN: Fast Vertical Mining of Sequential Generator Patterns ............ 476
  Philippe Fournier-Viger, Antonio Gomariz, Michal Šebek, and
  Martin Hlosta

Author Index ...................................................................................... 489