

2013 IEEE International Conference on Big Data

**Silicon Valley, California, USA
6 – 9 October 2013**

Pages 1-888



**IEEE Catalog Number: CFP13BGD-POD
ISBN: 978-1-4799-1294-0**

TABLE OF CONTENTS

Key Usage Patterns for Apache Hadoop in the Enterprise	1
<i>Amr Awadallah</i>	
The Berkeley Data Analytics Stack: Present and Future	2
<i>Mike Franklin</i>	
Using Crowdsourcing for Data Analytics	4
<i>Hector Garcia-Molina</i>	
Security – a Big Question for Big Data	5
<i>Roger Schell</i>	
On-line Learning Gossip Algorithm in Multi-agent Systems with Local Decision Rules	6
<i>Pascal Bianchi, Stephan Clemencon, Gemma Morral, J. Jakubowicz</i>	
Communication Efficient Algorithms for Fundamental Big Data Problems	15
<i>Peter Sanders, Sebastian Schlag, Ingo Muller</i>	
Map-based Graph Analysis on MapReduce	24
<i>Upa Gupta, Leonidas Fegaras</i>	
P-DOT: a Model of Computation for Big Data	31
<i>Tao Luo, Yin Liao, Guoliang Chen, Yunquan Zhang</i>	
Transparent Composite Model for Large Scale Image/video Processing	38
<i>En-Hui Yang, Xiang Yu</i>	
Elastic Algorithms for Guaranteeing Quality Monotonicity in Big Data Mining	45
<i>Rui Han, Lei Nie, Moustafa M. Ghanem, Yike Guo</i>	
HFSP: Size-based Scheduling for Hadoop	51
<i>Mario Pastorelli, Antonio Barbuzzi, Damiano Carra, Matteo Dell’Amico, Pietro Michiardi</i>	
An Evaluation Study of BigData Frameworks for Graph Processing	60
<i>Benedikt Elser, Alberto Montresor</i>	
Storing and Manipulating Environmental Big Data with JASMIN	68
<i>B. N. Lawrence, V. L. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pascoe, S. Pepler, M. Pritchard, A. Stephens</i>	
Efficient Gear-shifting for a Power-proportional Distributed Data-placement Method	76
<i>Hieu Hanh Le, Satoshi Hikida, Haruo Yokota</i>	
Agrios: a Hybrid Approach to Big Array Analytics	85
<i>Patrick Leyshock, David Maier, Kristin Tufte</i>	
Building a Generic Platform for Big Sensor Data Application	94
<i>Chun-Hsiang Lee, David Birch, Chao Wu, Dilshan Silva, Orestis Tsinalis, Yang Li, Shulin Yan, Moustafa Ghanem, Yike Guo</i>	
Locality-driven High-level I/O Aggregation for Processing Scientific Datasets	103
<i>Jialin Liu, Bradly Crysler, Yin Lu, Yong Chen</i>	
clusiVAT: a Mixed Visual/numerical Clustering Algorithm for Big Data	112
<i>Dheeraj Kumar, Marimuthu Palaniswami, Sutharshan Rajasegarar, Christopher Leckie, James C. Bezdek, Timothy C. Havens</i>	
Hardware Acceleration of Hadoop MapReduce	118
<i>Toshimori Honjo, Kazuki Oikawa</i>	
Optimizing the MapReduce Framework on Intel Xeon Phi Coprocessor	125
<i>Mian Lu, Lei Zhang, Huynh Phung Huynh, Zhongliang Ong, Yun Liang, Bingsheng He, Rick Siow Mong Goh, Richard Huynh</i>	
On the Performance and Energy Efficiency of Hadoop Deployment Models	131
<i>Eugen Feller, Lavanya Ramakrishnan, Christine Morin</i>	
Optimizing Throughput on Guaranteed-bandwidth WAN Networks for the Large Synoptic Survey Telescope (ISST)	137
<i>D. Michael Freemon</i>	
Feliss: Flexible Distributed Computing Framework with Light-weight Checkpointing	143
<i>Takuya Araki, Kazuyo Narita, Hiroshi Tamano</i>	
Algebraic Dataflows for Big Data Analysis	150
<i>Jonas Dias, Eduardo Ogasawara, Daniel De Oliveira, Fabio Porto, Patrick Valduriez, Marta Mattoso</i>	
Scalable and Robust Key Group Size Estimation for Reducer Load Balancing in MapReduce	156
<i>Wei Yan, Yuan Xue, Bradley Malin</i>	
Robot: an Efficient Model for Big Data Storage Systems Based on Erasure Coding	163
<i>Chao Yin, Jianzong Wang, Changsheng Xie, Jiguang Wan, Changlin Long, Wenjuan Bi</i>	

Multilevel Active Storage for Big Data Applications in High Performance Computing	169
<i>Chao Chen, Michael Lang, Yong Chen</i>	
GPU Accelerated Item-based Collaborative Filtering for Big-data Applications	175
<i>Chandima Hewa Nadungodage, Yuni Xia, John Jaehwan Lee, Myungcheol Lee, Choon Seo Park</i>	
GPU-accelerated Adaptive Compression Framework for Genomics Data	181
<i>Guixin Guo, Shuang Qiu, Zhiqiang Ye, BingQiang Wang, Lin Fang, Mian Lu, Simon See, Rui Mao</i>	
An Infrastructure for Automating Large-scale Performance Studies and Data Processing	187
<i>Deepal Jayasinghe, Josh Kimball, Tao Zhu, Siddharth Choudhary, Calton Pu</i>	
Kylin: an Efficient and Scalable Graph Data Processing System	193
<i>Li-Yung Ho, Tsung-Han Li, Jan-Jan Wu, Pangfeng Liu</i>	
Towards Hybrid Online On-demand Querying of Realtime Data with Stateful Complex Event Processing	199
<i>Quunzhi Zhou, Yogesh Simmhan, Viktor Prasanna</i>	
DDSN: Duplicate Detection to Reduce Both Storage and Bandwidth Consumption	206
<i>Jiaran Zhang, Xiaohui Yu, Yang Liu, Liwei Lin</i>	
A Reconfigurable Computing Architecture for Semantic Information Filtering	212
<i>Aalap Tripathy, Ka Chon Jeong, Atish Patra, Rabi Mahapatra</i>	
Iteration Aware Prefetching for Unstructured Grids	219
<i>Oyindamola O. Akande, Philip J. Rhodes</i>	
Measuring Inter-site Engagement	228
<i>Elad Yom-Tov, Mounia Lalmas, Ricardo Baeza-Yates, Georges Dupret, Janette Lehmann, Pinar Donmez</i>	
A Selective Checkpointing Mechanism for Query Plans in a Parallel Database System	237
<i>Ting Chen, Kenjiro Taura</i>	
CORE: Cross-object Redundancy for Efficient Data Repair in Storage Systems	246
<i>Kyumars Sheykh Esmaili, Lluís Pàmies-Juarez, Anwitaman Datta</i>	
H₂RDF+: High-performance Distributed Joins Over Large-scale RDF Graphs	255
<i>Nikolaos Papailiou, Ioannis Konstantinou, Dimitrios Tsoumakos, Panagiotis Karras, Nectarios Koziris</i>	
Direct QR Factorizations for Tall-and-skinny Matrices in Mapreduce Architectures	264
<i>Austin R. Benson, David F. Gleich, James Demmel</i>	
Adaptive File Management for Scientific Workflows on the Azure Cloud	273
<i>Radu Tudoran, Alexandru Costan, Ramin Rezaei Rad, Goetz Brasche, Gabriel Antoniu</i>	
Model-view Sensor Data Management in the Cloud	282
<i>Tian Guo, Thanasis G. Papaioannou, Karl Aberer</i>	
Spatio-temporal Indexing in Non-relational Distributed Databases	291
<i>Anthony Fox, Chris Eichelberger, James Hughes, Skylar Lyon</i>	
Scientific Discovery Through Weighted Sampling	300
<i>Lefteris Sidiropoulos, Martin Kersten, Peter Boncz</i>	
Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation	307
<i>Stefan Proll, Andreas Rauber</i>	
On the Use of Shared Storage in Shared-nothing Environments	313
<i>Krish K. R., Aleksandr Khasymski, Guanying Wang, Ali R. Butt, Gaurav Makkar</i>	
Self-adaptive Event Recognition for Intelligent Transport Management	319
<i>Alexander Artikis, Matthias Weidlich, Avigdor Gal, Vana Kalogeraki, Dimitrios Gunopulos</i>	
Improving Floating Point Compression Through Binary Masks	326
<i>Leonardo A. Bautista Gomez, Franck Cappello</i>	
Using Pattern-models to Guide SSD Deployment for Big Data Applications in HPC Systems	332
<i>Junjie Chen, Philip C. Roth, Yong Chen</i>	
Robust Crowdsourced Learning	338
<i>Zhiquan Liu, Luo Luo, Wu-Jun Li</i>	
Segmented Analysis for Reducing Data Movement	344
<i>Jialin Liu, Surendra Byna, Yong Chen</i>	
Continuous Hyperparameter Optimization for Large-scale Recommender Systems	350
<i>Simon Chan, Philip Treleaven, Licia Capra</i>	
4S: Scalable Subspace Search Scheme Overcoming Traditional Apriori Processing	359
<i>Hoang Vu Nguyen, Emmanuel Muller, Klemens Bohm</i>	
Computing Betweenness Centrality in External Memory	368
<i>Lars Arge, Michael T. Goodrich, Freek Van Walderveen</i>	
A Parallel Computing Platform for Training Large Scale Neural Networks	376
<i>Rong Gu, Furao Shen, Yihua Huang</i>	
Self-tuned Kernel Spectral Clustering for Large Scale Networks	385
<i>Raghvendra Mall, Rocco Langone, Johan A. K. Suykens</i>	

NUMA-optimized Parallel Breadth-first Search on Multicore Single-node System	394
<i>Yuichiro Yasui, Katsuki Fujisawa, Kazushige Goto</i>	
A Distributed Vertex-centric Approach for Pattern Matching in Massive Graphs	403
<i>Arash Fard, M. Usman Nisar, Lakshmish Ramaswamy, John A. Miller, Matthew Saltz</i>	
Fast Scalable Selection Algorithms for Large Scale Data	412
<i>Lee Parnell Thompson, Weijia Xu, Daniel P. Miranker</i>	
An NML-based Model Selection Criterion for General Relational Data Modeling	421
<i>Yoshiki Sakai, Kenji Yamanishi</i>	
Parallel Matrix Factorization for Binary Response	430
<i>Rajiv Khanna, Liang Zhang, Deepak Agarwal, Bee-Chung Chen</i>	
CallCab: a Unified Recommendation System for Carpooling and Regular Taxicab Services	439
<i>Desheng Zhang, Tian He, Yunhuai Liu, John A. Stankovic</i>	
Top-K Aggregation Over a Large Graph Using Shared-nothing Systems	448
<i>Abhirup Chakraborty</i>	
Distributed Confidence-weighted Classification on MapReduce	458
<i>Nemanja Djuric, Mihajlo Grbovic, Slobodan Vucetic</i>	
Scalable Context-aware Role Mining with MapReduce	467
<i>Zhiwei Yu, Raymond K. Wong, Chi-Hung Chi</i>	
Elver: Recommending Facebook Pages in Cold Start Situation Without Content Features	475
<i>Yusheng Xie, Zhengchang Chen, Kunpeng Zhang, Chen Jin, Yu Cheng, Ankit Agrawal, Alok Choudhary</i>	
Massively Scalable Near Duplicate Detection in Streams of Documents Using MDSH	480
<i>Paul Logasa Bogen, Christopher T. Symons, Amber McKenzie, Robert M. Patton, Robert E. Gillen</i>	
Incremental Algorithms for Closeness Centrality	487
<i>Ahmet Erdem Sariyuce, Kamer Kaya, Erik Saule, Umit V. Catalyurek</i>	
Classification of Big Velocity Data Via Cross-domain Canonical Correlation Analysis	493
<i>Bo Zhang, Zhong-Zhi Shi</i>	
A Distributed Tree Data Structure for Real-time OLAP on Cloud Architectures	499
<i>F. Dehne, Q. Kong, A. Rau-Chaplin, H. Zaboli, R. Zhou</i>	
DL-MPI: Enabling Data Locality Computation for MPI-based Data-intensive Applications	506
<i>Jiangling Yin, Andrew Foran, Jun Wang</i>	
Sparse Poisson Coding for High Dimensional Document Clustering	512
<i>Chenxia Wu, Haiqin Yang, Jianke Zhu, Jiemi Zhang, Irwin King, Michael R. Lyu</i>	
Fast OLAP Query Execution in Main Memory on Large Data in a Cluster	518
<i>Martin Weidner, Jonathan Dees, Peter Sanders</i>	
Group-scheme: SIMD-based Compression Algorithms for Web Text Data	525
<i>Xudong Zhang, Wayne Xin Zhao, Dongdong Shan, Hongfei Yan</i>	
Efficient Large Graph Pattern Mining for Big Data in the Cloud	531
<i>Chun-Chieh Chen, Kuan-Wei Lee, Chih-Chieh Chang, De-Nian Yang, Ming-Syan Chen</i>	
A Stream Partitioning Approach to Processing Large Scale Distributed Graph Datasets	537
<i>Rui Wang, Kenneth Chiu</i>	
Scalable Distributed Event Detection for Twitter	543
<i>Richard McCreadie, Craig Macdonald, Iadh Ounis, Miles Osborne, Sasa Petrovic</i>	
Analysis of GSM Calls Data for Understanding User Mobility Behavior	550
<i>Barbara Furletti, Lorenzo Gabrielli, Chiara Renso, Salvatore Rinzivillo</i>	
Scaling Concurrency of Personalized Semantic Search Over Large RDF Data	556
<i>Haizhou Fu, Hyeongsik Kim, Kemajor Anyanwu</i>	
A Hypergraph-partitioned Vertex Programming Approach for Large-scale Consensus Optimization	563
<i>Hui Miao, Xiangyang Liu, Bert Huang, Lise Getoor</i>	
A Higher-order Data Flow Model for Heterogeneous Big Data	569
<i>Simon Price, Peter A. Flach</i>	
Parallel Subgroup Discovery on Computing Clusters – First Results	575
<i>Daniel Trubold, Henrik Grosskreutz</i>	
DP-WHERE: Differentially Private Modeling of Human Mobility	580
<i>Darakhshan J. Mir, Sibren Isaacman, Ramon Caceres, Margaret Martonosi, Rebecca N. Wright</i>	
Malicious URL Filtering - a Big Data Application	589
<i>Min-Sheng Lin, Chien-Yi Chiu, Yuh-Jye Lee, Hsing-Kuo Pao</i>	
Zero-knowledge Private Graph Summarization	597
<i>Maryam Shoaran, Alex Thomo, Jens H. Weber-Jahnke</i>	
Scalable Network Traffic Visualization Using Compressed Graphs	606
<i>Lei Shi, Qi Liao, Xiaohua Sun, Yarui Chen, Chuang Lin</i>	
Breaking the Arc: Risk Control for Big Data	613
<i>Duncan Hodges, Sadie Creese</i>	

The BTWorld Use Case for Big Data Analytics: Description, MapReduce Logical Workflow, and Empirical Evaluation	622
<i>Tim Hegeman, Bogdan Ghit, Mihai Capota, Jan Hidders, Dick Epema, Alexandru Iosup</i>	
Modeling Heterogeneous Time Series Dynamics to Profile Big Sensor Data in Complex Physical Systems	631
<i>Bin Liu, Haijeng Chen, Abhishek Sharma, Guofei Jiang, Hui Xiong</i>	
Efficiently Extracting Frequent Subgraphs Using MapReduce	639
<i>Wei Lu, Gang Chen, Anthony K. H. Tung, Feng Zhao</i>	
Explaining the Product Range Effect in Purchase Data	648
<i>Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti</i>	
Large Scale Predictive Analytics for Real-time Energy Management	657
<i>Natasha Balac, Tamara Sipes, Nicole Wolter, Kenneth Nunes, Bob Sinkovits, Homa Karimabadi</i>	
Parallel Deterministic Annealing Clustering and Its Application to LC-MS Data Analysis	665
<i>Geoffrey Fox, D. R. Mani, Saumyadipta Pyne</i>	
Terabyte-scale Image Similarity Search: Experience and Best Practice	674
<i>Diana Moise, Denis Shestakov, Gylfi Gudmundsson, Laurent Amsaleg</i>	
HIG – an In-memory Database Platform Enabling Real-time Analyses of Genome Data	683
<i>Mathieu P. Schapranow, Hasso Plattner</i>	
Real-time Streaming Mobility Analytics	689
<i>Andras Garzo, Andras A. Benczur, Csaba Istvan Sidlo, Daniel Tahara, Erik Francis Wyatt</i>	
QuPARA: Query-driven Large-scale Portfolio Aggregate Risk Analysis on MapReduce	695
<i>A. Rau-Chaplin, B. Varghese, D. Wilson, Z. Yao, N. Zeh</i>	
Constructing Consumer Profiles from Social Media Data	702
<i>Mauricio Hernandez, Kirsten Hildrum, Prateek Jain, Rohit Wagle, Bogdan Alexe, Rajasekar Krishnamurthy, Ioana Roxana Stanoi, Chitra Venkatramani</i>	
CloudRS: an Error Correction Algorithm of High-throughput Sequencing Data Based on Scalable Framework	709
<i>Chien-Chih Chen, Yu-Jung Chang, Wei-Chun Chung, Der-Tsai Lee, Jan-Ming Ho</i>	
Demand Response Targeting Using Big Data Analytics	715
<i>Jungsuk Kwac, Ram Rajagopal</i>	
Building Dynamic Thermal Profiles of Energy Consumption for Individuals and Neighborhoods	723
<i>Adrian Albert, Ram Rajagopal</i>	
Terabyte-sized Image Computations on Hadoop Cluster Platforms	729
<i>Peter Bajcsy, Antoine Vandecreme, Julien Amelot, Phuong Nguyen, Joe Chalfoun, Mary Brady</i>	
A Fast and Scalable Method for Threat Detection in Large-scale DNS Logs	738
<i>Ron Begleiter, Yuval Elovici, Yona Hollander, Ori Mendelson, Lior Rokach, Roi Saltzman</i>	
Hourglass: a Library for Incremental Processing on Hadoop	742
<i>Matthew Hayes, Sam Shah</i>	
Correlation-based Performance Analysis for Full-system MapReduce Optimization	753
<i>Qi Guo, Yan Li, Tao Liu, Kun Wang, Guancheng Chen, Xiaoming Bao, Wentao Tang</i>	
Large Scale Ad Latency Analysis	762
<i>Mihajlo Grbovic, Jon Malkin, Hirakendu Das</i>	
Accelerating Semantic Graph Databases on Commodity Clusters	768
<i>Alessandro Morari, Vito Giovanni Castellana, David Haglin, John Feo, Jesse Weaver, Antonino Tumeo, Oreste Villa</i>	
Practical Distributed Classification Using the Alternating Direction Method of Multipliers Algorithm	773
<i>Peter Lubell-Doughtie, Jon Sondag</i>	
Scaling Deep Social Feeds at Pinterest	777
<i>Varun Sharma, Jeremy Carroll, Abhi Khune</i>	
Big Data Analytics on High Velocity Streams: a Case Study	784
<i>Thibaud Chardonens, Philippe Cudre-Mauroux, Martin Grund, Benoit Perroud</i>	
The Code Rebalancing Problem for a Storage-flexible Data Center Network	788
<i>Iryna Andriyanova, Alan Jule, Emina Soljanin</i>	
suvfs: a Virtual File System in Userspace That Supports Large Files	794
<i>Wasim Ahmad Bhat, S. M. K. Quadri</i>	
Distributed Storage Evaluation on a Three-wide Inter-data Center Deployment	799
<i>Yih-Farn Chen, Scott Daniels, Marios Hadjieleftheriou, Pingkai Liu, Chao Tian, Vinay Vaishampayan</i>	
Paired-replicas with Constant Repair Time: Loss Functions and Memorylessness	805
<i>Vinay Deolalikar</i>	
Efficient Updates in Cross-object Erasure-coded Storage Systems	810
<i>Kyumars Sheykh Esmaili, Aatish Chiniyah, Anwitaman Datta</i>	

Construction of Exact-BASIC Codes for Distributed Storage Systems at the MSR Point	815
<i>Hanxu Hou, Kenneth W. Shum, Hui Li</i>	
Minimum Storage Basic Codes: a System Perspective	821
<i>Xianxia Huang, Hui Li, Tai Zhou, Yumeng Zhang, Han Guo, Hanxu Hou, Huayu Zhang, Kai Lei</i>	
Layout-aware I/O Scheduling for Terabits Data Movement	826
<i>Youngjae Kim, Scott Atchley, Geoffroy R. Vallee, Galen M. Shipman</i>	
Reliability of Erasure Coded Storage Systems: a Geometric Approach	834
<i>Antonio Campello, Vinay A. Vaishampayan</i>	
Robustness of Emotion Extraction from 20th Century English Books	839
<i>Alberto Acerbi, Vasileios Lampos, R. Alexander Bentley</i>	
Back to Our Data — Experiments with NoSQL Technologies in the Humanities	847
<i>Tobias Blanke, Michael Bryant, Mark Hedges</i>	
Visualpage: Towards Large Scale Analysis of Nineteenth-century Print Culture	851
<i>Neal Audenaert, Natalie M. Houston</i>	
Visualization and Rhetoric: Key Concerns for Utilizing Big Data in Humanities Research: a Case Study of Vaccination Discourses: 1918-1919	859
<i>Kathleen Kerr, Bernice L. Hausman, Samah Gad, Waqas Javen</i>	
Humanities ‘Big Data’ - Myths, Challenges, and Lessons	867
<i>Amalia S. Levi</i>	
Digging Into Human Rights Violations: Data Modelling and Collective Memory	871
<i>B. Miller, A. Shrestha, J. Derby, J. Olive, K. Umapathy, F. Li, Y. Zhao</i>	
The Royal Birth of 2013: Analysing and Visualising Public Sentiment in the UK Using Twitter	880
<i>Vu Dung Nguyen, Blesson Vargheseb, Adam Barker</i>	
Bibliographic Records As Humanities Big Data	889
<i>Andrew Prescott</i>	
Customising Geoparsing and Georeferencing for Historical Texts	893
<i>C. J. Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, Patricia Murrieta-Flores</i>	
A Concept of Generic Workspace for Big Data Processing in Humanities	897
<i>Jedrzej Rybicki, Benedikt Von St. Vieth, Daniel Mallmann</i>	
From Assets to Stories Via the Google Cultural Institute Platform	905
<i>W. Brent Seales, Steve Crossan, Mark Yoshitake, Sertan Girgin</i>	
The Curious Identity of Michael Field and Its Implications for Humanities Research with the Semantic Web	911
<i>Susan Brown, John Simpson</i>	
Infectious Texts: Modeling Text Reuse in Nineteenth-century Newspapers	920
<i>David A. Smith, Ryan Cordell, Elizabeth Maddock Dillon</i>	
Mapping Mutable Genres in Structurally Complex Volumes	929
<i>Ted Underwood, Michael L. Black, Loretta Auvil, Boris Capitanu</i>	
CKM: a Shared Visual Analytical Tool for Large-scale Analysis of Audio-video Interviews	938
<i>Lu Xiao, Yan Luo, Steven High</i>	
A Case Study on Entity Resolution for Distant Processing of Big Humanities Data	947
<i>Weijia Xu, Maria Esteva, Jessica Trelogan, Todd Swinson</i>	
The Human Face of Crowdsourcing: a Citizen-led Crowdsourcing Case Study	955
<i>Sheryl Grant, Kristan E. Shawgo, Richard Marciano, Jeff Heard, Priscilla Ndiaye</i>	
Enterprise Pre-sales Forums: A Preliminary Study of Metadata and Content	959
<i>Vinay Deolalikar</i>	
A Cloud Service for the Evaluation of Company's Financial Health Using XBRL-based Financial Statements	963
<i>Wen-Chiao Hsu, Jyun-Yao Huang, Chi-Hao Chen, Chien-Yu Su, Hsiao-Chen Shih, Tzu-Ya Liao, I-En Liao</i>	
Real-time Data Analysis in ClowdFlows	968
<i>Janez Kranjc, Vid Podpecan, Nada Lavrac</i>	
Ma³tch: Privacy AND Knowledge ‘Dynamic Networked Collective Intelligence’	976
<i>Udo Kroon</i>	
Business Model Canvas Perspective on Big Data Applications	985
<i>F. Canan, Pembe Muhtaroglu, Seniz Demir, Murat Obali, Canan Girgin</i>	
Advancing Value Creation and Value Capture in Data-intensive Contexts	991
<i>Roman Ferrando-Llopis, David Lopez-Berzosa, Catherine Mulligan</i>	
OpenFridge: a Platform for Data Economy for Energy Efficiency Data	996
<i>Slobodanka Dana Kathrin Tomic, Anna Fensel</i>	

A Study of Innovation Network Database Construction by Using Big Data and an Enterprise Strategy Model	1001
<i>Wen Zhou, Shu-Tao Ye, Xiao-Long Lu</i>	
Enhanced User Data Privacy with Pay-by-data Model	1006
<i>Chao Wu, Yike Guo</i>	
Query Optimization Over a Heterogeneously Distributed Scientific Database	1011
<i>Helen X. Xiang</i>	
Enterprise Data Economy: A Hadoop-driven Model and Strategy	1018
<i>Wuheng Luo</i>	
Understanding the Value of (Big) Data	1024
<i>Koutroumpis Pantelis, Leiponen Aija</i>	
Hash in a Flash: Hash Tables for Flash Devices	1029
<i>Tyler Clemons, S. M. Faisal, Shirish Tatikonda, Charu Aggarwal, Srinivasan Parthasarathy</i>	
Memory System Characterization of Big Data Workloads	1037
<i>Martin Dimitrov, Karthik Kumar, Patrick Lu, Vish Viswanathan, Thomas Willhalm</i>	
Optimizing a MapReduce Module of Preprocessing High-throughput DNA Sequencing Data	1045
<i>Wei-Chun Chung, Yu-Jung Chang, Chien-Chih Chen, Der-Tsai Lee, Jan-Ming Ho</i>	
Performance Evaluation of R with Intel Xeon Phi Coprocessor	1051
<i>Yaakoub El-Khamra, Niall Gaffney, David Walling, Eric Wernert, Weijia Xu, Hui Zhang</i>	
A Performance Evaluation of Hive for Scientific Data Management	1059
<i>Taoying Liu, Jing Liu, Hong Liu, Wei Li</i>	
Evaluating Task Scheduling in Hadoop-based Cloud Systems	1067
<i>Shengyuan Liu, Jungang Xu, Zongzhen Liu, Xu Liu</i>	
Efficient Near-duplicate Document Detection Using FPGAs	1074
<i>Xi Luo, Walid Najjar, Vagelis Hristidis</i>	
Workload-aware Aggregate Maintenance in Columnar In-Memory Databases	1082
<i>Stephan Muller, Lars Butzmann, Stefan Klauck, Hasso Plattner</i>	
Virtualization I/O Optimization Based on Shared Memory	1090
<i>Fengfeng Ning, Chuliang Weng, Yuan Luo</i>	
An Ensemble MIC-based Approach for Performance Diagnosis in Big Data Platform	1098
<i>Pengfei Chen, Yong Qi, Xinyi Li, Li Su</i>	
A Reconfigurable Stream Compression Hardware Based on Static Symbol-lookup Table	1106
<i>Shinichi Yamagiwa, Hiroshi Sakamoto</i>	
NativeTask: a Hadoop Compatible Framework for High Performance	1114
<i>Dong Yang, Xiang Zhong, Dong Yan, Fangqin Dai, Xusen Yin, Cheng Lian, Zhongliang Zhu, Weihua Jiang, Gansha Wu</i>	
On Mixing High-speed Updates and In-memory Queries: a Big-Data Architecture for Real-time Analytics	1122
<i>Tao Zhong, Kshitij A. Doshi, Xi Tang, Ting Lou, Zhongyan Lu, Hong Li</i>	
AxPUE: Application Level Metrics for Power Usage Effectiveness in Data Centers	1130
<i>Runlin Zhou, Yingjie Shi, Chungze Zhu</i>	
The Implications from Benchmarking Three Big Data Systems	1138
<i>Jing Quan, Yingjie Shi, Ming Zhao, Wei Yang</i>	
A Characterization of Big Data Benchmarks	1146
<i>Wen Xiong, Zhibin Yu, Zhendong Bei, Juanjuan Zhao, Fan Zhang, Yubin Zou, Xue Bai, Ye Li, Chengzhong Xu</i>	
Dynamic Reduction of Query Result Sets for Interactive Visualizatoin	1154
<i>Leilani Battle, Michael Stonebraker, Remco Chang</i>	
Typograph: Multiscale Spatial Exploration of Text Documents	1162
<i>Alex Endert, Russ Burtner, Nick Cramer, Ralph Perko, Shawn Hampton, Kristin Cook</i>	
VisReduce: Fast and Responsive Incremental Information Visualization of Large Datasets	1170
<i>Jean-Francois Im, Felix Giguere Villegas, Michael J. McGuffin</i>	
A System for Large-scale Visualization of Streaming Doppler Data	1178
<i>Peter Kristof, Bedrich Benes, Carol X. Song, Lan Zhao</i>	
Overplotting: Unified Solutions Under Abstract Rendering	1186
<i>Joseph Cottam, Andrew Lumsdaine, Peter Wang</i>	
Visualization of Streaming Data: Observing Change and Context in Information Visualization Techniques	1194
<i>Milos Krstajic, Daniel A. Keim</i>	
CompactMap: a Mental Map Preserving Visual Interface for Streaming Text Data	1201
<i>Xiaotong Liu, Yifan Hu, Stephen North, Han-Wei Shen</i>	
Egocentric Storylines for Visual Analysis of Large Dynamic Graphs	1209
<i>Chris W. Muelder, Tarik Crnovrsanin, Arnaud Sallaberry, Kwan-Liu Ma</i>	

GPU-Accelerated Incremental Correlation Clustering of Large Data with Visual Feedback	1216
<i>Eric Papenhausen, Bing Wang, Sungsoo Ha, Alla Zelenyuk, Dan Imre, Klaus Mueller</i>	
Visualization of Big SPH Simulations Via Compressed Octree Grids	1224
<i>Florian Reichl, Marc Treib, Rudiger Westermann</i>	
A Novel Visual Analytics Approach for Clustering Large-scale Social Data	1232
<i>Zhangye Wang, Juanxia Zhou, Wei Chen, Chang Chen, Jiyuan Liao, Ross Maciejewski</i>	
DriveSense: Contextual Handling of Large-scale Route Map Data for the Automobile	1240
<i>Frederik Wiehr, Vidya Setlur, Alark Joshi</i>	
A Big Data Analytics Framework for Scientific Data Management	1248
<i>Sandro Fiore, Cosimo Palazzo, Alessandro D'Anca, Ian Foster, Dean N. Williams, Giovanni Aloisio</i>	
Searching Inter-disciplinary Scientific Big Data Based on Latent Correlation Analysis	1256
<i>Eloy Gonzales, Bun Theang Ong, Koji Zettsu</i>	
Complete Storm Identification Algorithms from Big Raw Rainfall Data Using MapReduce Framework	1260
<i>Kulsawasd Jitkajornwanich, Upa Gupta, Sakthi Kumaran Shanmuganathan, Ramez Elmasri, Leonidas Fegaras, John McEnery</i>	
A Scalable Data Analysis Platform for Metagenomics	1268
<i>Wei Tang, Jared Wilkening, Narayan Desai, Wolfgang Gerlach, Andreas Wilke, Folker Meyer</i>	
Rethinking Data Management for Big Data Scientific Workflows	1274
<i>Karan Vahi, Mats Rynge, Gideon Juve, Rajiv Mayani, Ewa Deelman</i>	
SciFlow: a Dataflow-driven Model Architecture for Scientific Computing Using Hadoop	1283
<i>Pengfei Xuan, Yueli Zheng, Sapna Sarupria, Amy Apon</i>	
Assessment of Dimensionality Reduction Based on Communication Channel Model; Application to Immersive Information Visualization	1292
<i>Mohammadreza Babae, Mihai Datcu, Gerhard Rigoll</i>	
Hierarchical Feature Learning from Sensorial Data by Spherical Clustering	1298
<i>Bonny Banerjee, Jayanta K. Dutta</i>	
Efficient Learning from Explanation of Prediction Errors in Streaming Data	1305
<i>Bonny Banerjee, Jayanta K. Dutta</i>	
Distributed Pivot Clustering with Arbitrary Distance Functions	1312
<i>L. Karl Branting</i>	
Nearest Neighbor Classification Using Bottom-k Sketches	1319
<i>Søren Dahlgaard, Christian Igel, Mikkel Thorup</i>	
Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets	1326
<i>Ciro Donalek, S. G. Djorgovski, Ashish A. Mahabal, Matthew J. Graham, Andrew J. Drake, Arun Kumar, N. Sajeeth Philip, Thomas J. Fuchs, Michael J. Turmon, Michael Ting-Chang Yang, Giuseppe Longo</i>	
How Data Partitioning Strategies and Subset Size Influence the Performance of an Ensemble?	1333
<i>Majed Farrash, Wenjia Wang</i>	
Fast Change Point Detection for Electricity Market Analysis	1341
<i>William Gu, Jaesik Choi, Ming Gu, Horst Simon, Kesheng Wu</i>	
A Novel Integrated Method for Human Multiplex Protein Subcellular Localization Prediction	1349
<i>Hong Gu, Junzhe Cao</i>	
Learning from Multiple Data Sets with Different Missing Attributes and Privacy Policies: Parallel Distributed Fuzzy Genetics-based Machine Learning Approach	1354
<i>Hisao Ishibuchi, Masakazu Yamane, Yusuke Nojima</i>	
Data Chaos: an Entropy Based Mapreduce Framework for Scalable Learning	1362
<i>Jiaoyan Chen, Huajun Chen, Xi Chen, Guozhou Zheng, Zhaohui Wu</i>	
Exploring Sketches for Probability Estimation with Sublinear Memory	1370
<i>Anthony Kleerekoper, Mikel Lujan, Gavin Brown</i>	
Agglomerative Co-clustering for Synonymous Phrases Based on Common Effects and Influences	1378
<i>Koji Kumanami, Kazuhiro Seki, Kuniaki Uehara</i>	
Leveraging Memory Mapping for Fast and Scalable Graph Computation on a PC	1386
<i>Zhiyuan Lin, Duen Horng (Polo) Chau, U. Kang</i>	
Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier	1390
<i>Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen, Genshe Chen</i>	
Meta-learning for Large Scale Machine Learning with MapReduce	1396
<i>Xuan Liu, Xiaoguang Wang, Stan Matwin, Nathalie Japkowicz</i>	
Frequent Itemset Mining for Big Data	1402
<i>Sandy Moens, Emin Aksehirli, Bart Goethals</i>	
Evaluating Parallel Logistic Regression Models	1410
<i>Haoruo Peng, Ding Liang, Cyrus Choi</i>	

Approximate Triangle Counting Algorithms on Multi-cores	1418
<i>Mahmudur Rahman, Mohammad Al Hasan</i>	
Tree Labeled LDA: a Hierarchical Model for Web Summaries	1425
<i>Anton Slutsky, Xiaohua Hu, Yuan An</i>	
Nearest Neighbour Regression Outperforms Model-based Prediction of Specific Star Formation Rate	1432
<i>Kristoffer Stensbo-Smidt, Christian Igel, Andrew Zirm, Kim Steenstrup Pedersen</i>	
MapReduce Implementation of Variational Bayesian Probabilistic Matrix Factorization Algorithm	1436
<i>Naveen C. Tewari, Hari M. Koduvely, Sarbendu Guha, Arun Yadav, Gladbin David</i>	
A Unified Framework for Predicting Attributes and Links in Social Networks	1444
<i>Xusen Yin, Bin Wu, Xiuqin Lin</i>	
Scalable Approximation of Kernel Fuzzy C-means	1452
<i>Zijian Zhang, Timothy C. Havens</i>	
Large-scale Restricted Boltzmann Machines on Single GPU	1460
<i>Yun Zhu, Yanqing Zhang, Yi Pan</i>	
Lung Transplant Outcome Prediction Using UNOS Data	1466
<i>Ankit Agrawal, Reda Al-Bahrani, Mark J. Russo, Jaishankar Raman, Alok Choudhary</i>	
Colon Cancer Survival Prediction Using Ensemble Data Mining on SEER Data	1474
<i>Reda Al-Bahrani, Ankit Agrawal, Alok Choudhary</i>	
A Look at Challenges and Opportunities of Big Data Analytics in Healthcare	1482
<i>Raghunath Nambiar, Ruchie Bhardwaj, Adhiraaj Sethi, Rajesh Vargheese</i>	
Multidimensional Analysis of Fetal Growth Curves	1488
<i>Mario A. Bochicchio, Antonella Longo, Lucia Vaira, Antonio Malvasi, Andrea Tinelli</i>	
OWL Reasoning Over Big Biomedical Data	1494
<i>Xi Chen, Huajun Chen, Ningyu Zhang, Jiaoyan Chen, Zhaohui Wu</i>	
KUChemBio: a Repository of Computational Chemical Biology Data Sets	1502
<i>Aaron Smalter Hall, Jun Huan</i>	
Parallel and Memory-efficient Burrows-Wheeler Transform	1508
<i>Shinya Hayashi, Kenjiro Taura</i>	
Content-based Assessment of the Credibility of Online Healthcare Information	1516
<i>Meeyoung Park, Hariprasad Sampathkumar, Bo Luo, Xue-Wen Chen</i>	
BIG DATA Infrastructures for Pharmaceutical Research	1524
<i>Christian Seebode, Matthias Ort, Christian Regenbrecht, Martin Peuker</i>	
Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients	1529
<i>Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, Brian Muckian</i>	
The Microsoft Academic Search Challenges at KDD Cup 2013	1537
<i>Martine De Cock, Senjuti Basu Roy, Swapna Savvana, Vani Mandava, Brian Dalessandro, Claudia Perlich, William Cukierski, Ben Hamner</i>	
Bibliometric-enhanced Retrieval Models for Big Scholarly Information Systems	1541
<i>Philipp Mayr, Peter Mutschke</i>	
Academic Publishing As a Social Media Paradigm	1545
<i>Michael E. Payne, Linh B. Ngo, Amy W. Apon</i>	
Big Spatial Data Mining	1549
<i>Shuliang Wang, Gangyi Ding, Ming Zhong</i>	
Modeling and Querying Data in NoSQL Databases	1558
<i>Karamjit Kaur, Rinkle Rani</i>	
Elastic Data Partitioning for Cloud-based SQL Processing Systems	1565
<i>Lipyeow Lim</i>	
Parallel SECOND: Practical and Efficient Mobility Data Processing in the Cloud	1574
<i>Jiamin Lu, Ralf Hartmut Guting</i>	
Index-based Join Operations in Hive	1583
<i>Mahsa Mofidpoor, Nematollaah Shiri, T. Radhakrishnan</i>	
SLA Data Management Criteria	1591
<i>Katerina Stamou, Verena Kantere, Jean-Henry Morin</i>	
Fast Solution of Load Shedding Problems Via a Sequence of Linear Programs	1600
<i>Harish S. Bhat, Garnet J. Vaz, Juan C. Meza</i>	
Alarm Prediction in Large-scale Sensor Networks - a Case Study in Railroad	1606
<i>Hongfei Li, Buyue Qian, Dhaivat Parikh, Arun Hampapur</i>	
MISTRAL: an Architecture for Low-latency Analytics on Massive Time Series	1614
<i>Alice Marascu, Pascal Pompey, Eric Bouillet, Olivier Verscheure, Michael Wurst, Martin Grund, Philippe Cudre-Mauroux</i>	
Yellow Cabs As Red Corpuscles	1621
<i>Timothy H. Savage, Huy T. Vo</i>	

Scalable Prediction of Energy Consumption Using Incremental Time Series Clustering	1628
<i>Yogesh Simmhan, Muhammad Usman Noor</i>	
A Big Data Driven Model for Taxi Drivers' Airport Pick-up Decisions in New York City	1636
<i>M. Anil Yazici, Camille Kamga, Abhishek Singhal</i>	
Managing Massive Graphs in Relational DBMS	1644
<i>Ruiwen Chen</i>	
A Distributed Approach for Graph-Oriented Multidimensional Analysis	1652
<i>Benoît Denis, Amine Ghrab, Sabri Skhiri</i>	
Constructing E-Tourism Platform Based on Service Value Broker: a Knowledge Management Perspective	1660
<i>Yucong Duan, Yongzhi Wang, Jinpeng Wei, Ajay Kattepur, Wencai Du</i>	
ADraw: a Novel Social Network Visualization Tool with Attribute-based Layout and Coloring	1668
<i>Zhenwen Wang, Weidong Xiao, Bin Ge, Hao Xu</i>	
IntegrityMR: Integrity Assurance Framework for Big Data Analytics and Management Applications	1676
<i>Yongzhi Wang, Jinpeng Wei, Mudhakar Srivatsa, Yucong Duan, Wencai Du</i>	
Local Join Optimization Over a Heterogeneously Distributed Scientific Database	1684
<i>Helen X. Xiang</i>	
Core-based Community Evolution in Mobile Social Networks	1689
<i>Hao Xu, Weidong Xiao, Daquan Tang, Jiuyang Tang, Zhenwen Wang</i>	
Super-sequence Frequent Pattern Mining on Sequential Dataset	1695
<i>Xinran Yu, Turgay Korkmaz</i>	
Exploring Big Data in Small Forms: a Multi-layered Knowledge Extraction of Social Networks	1703
<i>Yun Wei Zhao, Willem-Jan Van Den Heuvel, Xiaojun Ye</i>	
Provenance Comparison for Large-scale Knowledge Discovery	1711
<i>Xiang Zhao, Bin Ge, Jiuyang Tang, Weidong Xiao, Haichuan Shang</i>	
Re-projection of Terabyte-Sized Images	1719
<i>Peter Bajcsy, Antoine Vandecreme, Mary Brady</i>	
Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis	1720
<i>Daniel Cheng, Peter Schretten, Nathan Kronenfeld, Neil Bozowsky, William Wright</i>	
Optimizing Queries Over Semantically Integrated Datasets on MapReduce Platforms	1723
<i>Hyeongsik Kim, Kemajor Anyanwu</i>	
Secure Decoupled Linkage (SDLink) System for Building a Social Genome	1725
<i>Hye-Chung Kum, Ashok Krishnamurthy, Darshana Pathak, Michael K. Reiter, Stanley Ahalt</i>	
Risk Adjustment of Patient Expenditures: a Big Data Analytics Approach	1730
<i>Lin Li, Saeed Bagheri, Helena Goote, Asif Hasan, Gregg Hazard</i>	
Parallel Auto-encoder for Efficient Outlier Detection	1733
<i>Yunlong Ma, Peng Zhang, Yanan Cao, Li Guo</i>	
New Factors for Identifying Influential Bloggers	1736
<i>Teng-Sheng Moh, SivaNaga Prasad Shola</i>	
A Scalable Infrastructure of Interactive Evolutionary Computation to Evolve Services Online with Data	1746
<i>Masaharu Munetomo, Shintaro Bando</i>	
Big Data for Business Managers - Bridging the Gap Between Potential and Value	1747
<i>Anmol Rajpurohit</i>	
Granularity-based Temporal Data Mining in Hospital Information System	1750
<i>Shusaku Tsumoto, Shoji Hirano, Haruko Iwata</i>	
Observation of Matthew Effects in Sina Weibo Microblogger	1759
<i>Mengmeng Yang, Yi Zhou, Qu Zhou, Kai Chen, Jianhua He, Xiaokang Yang</i>	
A Framework of Spatial Co-location Mining on MapReduce	1762
<i>Jin Soung Yoo, Douglas Boulware</i>	
Access Control for Big Data Using Data Content	1763
<i>Wenrong Zeng, Yuhao Yang, Bo Luo</i>	
Knowledge Cubes - A Proposal for Scalable and Semantically-guided Management of Big Data	1766
<i>Amgad Madkour, Walid G. Aref, Saleh Basalamah</i>	
Author Index	