

2013 IEEE International Congress on Big Data

(BigData Congress 2013)

**Formerly Known as International Conference on Services
Economics (SE)**

**Santa Clara, California, USA
27 June – 2 July 2013**



**IEEE Catalog Number: CFP13SEV-POD
ISBN: 978-1-4799-0182-1**

2013 IEEE International Congress on Big Data

BigData Congress 2013

Table of Contents

Message from the General and Program Chairs.....	xi
Organizing Committee.....	xii
Program Committee.....	xv
Support Team.....	xvii
IEEE Computer Society Technical Committee on Services Computing.....	xviii

Research Tracks

Research Session 1 — Big Data Scalability

A Database-Hadoop Hybrid Approach to Scalable Machine Learning	1
<i>Makoto Yui and Isao Kojima</i>	
Performance Overhead among Three Hypervisors: An Experimental Study Using Hadoop Benchmarks	9
<i>Jack Li, Qingyang Wang, Deepal Jayasinghe, Junhee Park, Tao Zhu, and Calton Pu</i>	
Data Allocation in Scalable Distributed Database Systems Based on Time Series Forecasting	17
<i>Shun Pun Li and Man Hon Wong</i>	

Research Session 2 — Privacy Issues

A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case	25
<i>Omar Hasan, Benjamin Habegger, Lionel Brunie, Nadia Bennani, and Ernesto Damiani</i>	
Approximate Two-Party Privacy-Preserving String Matching with Linear Complexity	31
<i>Martin Beck and Florian Kerschbaum</i>	
Engineering Privacy for Big Data Apps with the Unified Modeling Language	38
<i>Dawn N. Jutla, Peter Bodorik, and Sohail Ali</i>	

Research Session 3 — Big Data Mining

Milieu: Lightweight and Configurable Big Data Provenance for Science	46
<i>You-Wei Cheah, Richard Canon, Beth Plale, and Lavanya Ramakrishnan</i>	
Consistent Process Mining over Big Data Triple Stores	54
<i>Antonia Azzini and Paolo Ceravolo</i>	

Research Session 4 — Big Data Cloud

Towards Cloud-Based Analytics-as-a-Service (CLAAaaS) for Big Data Analytics in the Cloud	62
<i>Farhana Zulkernine, Patrick Martin, Ying Zou, Michael Bauer, Femida Gwadry-Sridhar, and Ashraf Aboulnaga</i>	
Scalable and Trustworthy Cross-Enterprise WfMSs by Cloud Collaboration	70
<i>Gwan-Hwan Hwang, Yi-Chan Kao, and Yu-Cheng Hsiao</i>	
A Bandwidth-Conscious Caching Scheme for Mobile Devices	78
<i>Badari N. Thyamagondlu, Victor W. Chu, and Raymond K. Wong</i>	

Research Session 5 — Pervasive Big Data

Towards a Quality-centric Big Data Architecture for Federated Sensor Services	86
<i>Lakshmish Ramaswamy, Victor Lawson, and Siva Venkat Gogineni</i>	
Learning Classifiers from Chains of Multiple Interlinked RDF Data Stores	94
<i>Harris T. Lin and Vasant Honavar</i>	
Multi-resolution Social Network Community Identification and Maintenance on Big Data Platform	102
<i>Hidayet Aksu, Mustafa Canim, Yuan-Chi Chang, Ibrahim Korpeoglu, and Özgür Ulusoy</i>	

Research Session 6 — Rule Mining

Online Association Rule Mining over Fast Data	110
<i>Erdi Ölmezogullari and Ismail Ari</i>	
Approximate Incremental Big-Data Harmonization	118
<i>Puneet Agarwal, Gautam Shroff, and Pankaj Malhotra</i>	
Countering the Concept-Drift Problem in Big Data Using iOVFDT	126
<i>Hang Yang and Simon Fong</i>	

Research Special Session 1 — Social Media Data Analytics

MapReduce-Based SimRank Computation and Its Application in Social Recommender System	133
<i>Lina Li, Cuiping Li, Hong Chen, and Xiaoyong Du</i>	
Duplicate Detection for Identifying Social Spam in Microblogs	141
<i>Qunyan Zhang, Haixin Ma, Weining Qian, and Aoying Zhou</i>	
Graph-Based Hierarchical Categorization of Microblog Users	149
<i>Kun Yue, Minqi Zhou, Jixian Zhang, Ping Zhang, Qiyu Fang, and Weiyi Liu</i>	

Research Special Session 2 — Big Data Querying

Scalable Parallel Join for Huge Tables	157
<i>Nianlong Weng, Minqi Zhou, Ming-Chien Shan, and Aoying Zhou</i>	
Efficient SPARQL Query Evaluation in a Database Cluster	165
<i>Fang Du, Haoqiong Bian, Yueguo Chen, and Xiaoyong Du</i>	
A BSP-Based Parallel Iterative Processing System with Multiple Partition Strategies for Big Graphs	173
<i>Zhigang Wang, Yubin Bao, Yu Gu, Fangling Leng, Ge Yu, Chao Deng, and Leitao Guo</i>	

Research Special Session 3 — Big Graph Data

Fast Similar Subgraph Search with Maximum Common Connected Subgraph Constraints	181
<i>Huiqi Hu, Guoliang Li, and Jianhua Feng</i>	
Revealing the Causes of Dynamic Change in Protein-Protein Interaction Network	189
<i>Yang Guo, Xuequn Shang, Jing Li, and Zhanhuai Li</i>	
SPTI: Efficient Answering the Shortest Path Query on Large Graphs	195
<i>Yifei Zhang and Guoren Wang</i>	

Research Special Session 4 — MapReduce

Efficient Probabilistic Skyline Query Processing in MapReduce	203
<i>Linlin Ding, Guoren Wang, Junchang Xin, and Ye Yuan</i>	
A Throughput Driven Task Scheduler for Improving MapReduce Performance in Job-Intensive Environments	211
<i>Xite Wang, Derong Shen, Ge Yu, Tiezheng Nie, and Yue Kou</i>	

Research Special Session 5 — Big Data Infrastructure

Massive Parallel Join in NUMA Architecture	219
<i>Wei He, Minqi Zhou, Xueqing Gong, and Xiaofeng He</i>	
Optimal Self-Healing of Service-Oriented Systems with Incomplete Information	227
<i>Hongbing Wang, Xiaojun Wang, and Qi Yu</i>	

Applications and Industry Tracks

Industry and Application Session 1 — Privacy Protection

Challenges of Privacy Protection in Big Data Analytics	235
<i>Meiko Jensen</i>	
Decentralized Trust Driven Access Control for Mobile Content Sharing	239
<i>B.S. Vidyalakshmi, Raymond K. Wong, and Chi-Hung Chi</i>	

Industry and Application Session 2 — Big Data Analysis

Cost and Time Aware Ant Colony Algorithm for Data Replica in Alpha Magnetic Spectrometer Experiment	247
<i>Lijuan Wang, Junzhou Luo, Jun Shen, and Fang Dong</i>	
Techniques for Graph Analytics on Big Data	255
<i>M. Usman Nisar, Arash Fard, and John A. Miller</i>	
Data Abstraction and Visualisation in Next Step: Experiences from a Government Services Delivery Trial	263
<i>Sanat Kumar Bista, Surya Nepal, and Cecile Paris</i>	

Industry and Application Session 3 — System Architecture

CroTIS-Crowdsourcing Based Traffic Information System	271
<i>T. Roopa, Anantharaman Narayana Iyer, and Shanta Rangaswamy</i>	
Characterization of 3G Data-Plane Traffic and Application towards Centralized Control and Management for Software Defined Networking	278
<i>Long Qian, Bin Wu, Renbo Zhang, Weiyu Zhang, and Min Luo</i>	

Industry and Application Session 4 — Domain Knowledge

Smart M2M Data Filtering Using Domain-Specific Thresholds in Domain-Agnostic Platforms	286
<i>Apostolos Papageorgiou, Mischa Schmidt, Jaeseung Song, and Nobuharu Kami</i>	
Full Recognition of Massive Products Based on Property Set	294
<i>Li Kuang, Liang Chen, Yanan Xie, and Jian Wu</i>	
Learning Classifiers from Distributional Data	302
<i>Harris T. Lin, Sanghack Lee, Ngot Bui, and Vasant Honavar</i>	

Industry and Application Session 5 — Workflow Computing

Highly Scalable Sequential Pattern Mining Based on MapReduce Model on the Cloud	310
<i>Chun-Chieh Chen, Chi-Yao Tseng, and Ming-Syan Chen</i>	
Small Is Beautiful: Summarizing Scientific Workflows Using Semantic Annotations	318
<i>Pinar Alper, Khalid Belhajjame, Carole Goble, and Pinar Karagoz</i>	

Industry and Application Session 6 — Load Balancing and Testing

SPOAN: Load Balancing Replica Placement Strategy for Large Scale Biometric Identification Service	326
<i>Takatoshi Kitano and Leiming Su</i>	
Benchmarking Apache Accumulo BigData Distributed Table Store Using Its Continuous Test Suite	334
<i>Ranjan Sen, Andrew Farris, and Peter Guerra</i>	
Surveying Systems of Global Climate Change Simulation Data and Access Logs to Them	342
<i>Toshihiro Nemoto and Masaru Kitsuregawa</i>	

Industry and Application Session 7 —Big Data Algorithm

Fast Quasi-biclique Mining with Giraph	347
<i>Hsiao-Fei Liu, Chung-Tsai Su, and An-Chiang Chu</i>	
Economical Data-Intensive Service Provision Supported with a Modified Genetic Algorithm	355
<i>Lijuan Wang and Jun Shen</i>	
Data Mining Approaches for Packaging Yield Prediction in the Post-fabrication Process	363
<i>Seung Hwan Park, Cheong-Sool Park, Jun Seok Kim, Sung-Shick Kim, Jun-Geol Baek, and Daewoong An</i>	

Industry and Application Session 8 — Big Data Querying

Distributed SPARQL Query Answering over RDF Data Streams	369
<i>Marcello Leida and Andrej Chu</i>	
Point and Interval Estimation Method for Auto-regressive Model with Nonnormal Error	379
<i>Bo Mi Lim, Jongwoo Kim, Sung-Shick Kim, and Jun-Geol Baek</i>	
Labeling Instances in Evolving Data Streams with MapReduce	387
<i>Ahsanul Haque, Brandon Parker, and Latifur Khan</i>	

Industry and Application Session 9 — Big Data Tool

RSenter: Tool for Topics and Terms Extraction from Unstructured Data Debris	395
<i>Richard K. Lomotey and Ralph Deters</i>	
Service-Generated Big Data and Big Data-as-a-Service: An Overview	403
<i>Zibin Zheng, Jieming Zhu, and Michael R. Lyu</i>	

Work-in-Progress Tracks

WIP Session 1 — Context Awareness Model

Big Data Infrastructure for Active Situation Awareness on Social Network Services	411
<i>Incheon Paik, Takazumi Tanaka, Hiroki Ohashi, and Wuhui Chen</i>	
Fraud Detection on Large Scale Social Networks	413
<i>Yaya Sylla and Pierre Morizet-Mahoudeaux</i>	
Network-Aware Web Service Selection in Dynamic Environment	415
<i>Lei Yu, Wang Zhili, Shao Mingfang, Wang Zishuo, Luoming Meng, and Qiu Xue-Song</i>	
Distributed Processing from Large Scale Sensor Network Using Hadoop	417
<i>Rosangela de Fátima Pereira, Marcelo Risse de Andrade, Artur Carvalho Zucchi, Karen Langona, Walter Akio Goya, Nelson Mimura Gonzalez, Tereza Cristina Melo Brito de Carvalho, Jan-Erik Mångs, and Azimeh Sefidcon</i>	
Analysis of Technology Trends Based on Big Data	419
<i>Aviv Segev, Chihoon Jung, and Sukhwan Jung</i>	

WIP Session 2 — Big Data Trends

Storage Mining: Where IT Management Meets Big Data Analytics	421
<i>Yang Song, Gabriel Alatorre, Nagapramod Mandagere, and Aameek Singh</i>	
RABID—A General Distributed R Processing Framework Targeting Large Data-Set Problems	423
<i>Hao Lin, Shuo Yang, and Samuel P. Midkiff</i>	
Distributed Stochastic Aware Random Forests—Efficient Data Mining for Big Data	425
<i>Joaquim Assunção, Paulo Fernandes, Lucelene Lopes, and Silvio Normey</i>	
Data for All: A Systems Approach to Accelerate the Path from Data to Insight	427
<i>Eser Kandogan, Mary Roth, Cheryl Kieliszewski, Fatma Özcan, Bob Schloss, and Marc-Thomas Schmidt</i>	

WIP Session 3 — Big Data Management

The Knowledge Service Project in the Era of Big Data	429
<i>Dongfeng Cai, Yu Bai, Guiping Zhang, and Fang Cai</i>	
Secure Outsourcing of Network Flow Data Analysis	431
<i>Mohamed Nassar, Bechara al Bouna, and Qutaibah Malluhi</i>	
Graph Data Warehouse: Steps to Integrating Graph Databases Into the Traditional Conceptual Structure of a Data Warehouse	433
<i>Yunkai Liu and Theresa M. Vitolo</i>	
IP2User—Identifying the Username of an IP Address in Network-Related Events	435
<i>Asaf Shabtai, Idan Morad, Eyal Kolman, Erel Eran, Alex Vaystikh, Eyal Gruss, Lior Rokach, and Yuval Elovici</i>	
Cloud Distributed Processing Using Trade Wind	437
<i>Nelson Mimura Gonzalez, Walter Akio Goya, Karen Langona, Rosangela de Fátima Pereira, Marcelo Risse de Andrade, Artur Carvalho Zucchi, Tereza Cristina Melo Brito de Carvalho, Jan-Erik Mångs, and Azimeh Sefidcon</i>	
Effective Interpretation of Bucket Testing Results through Big Data Analytics	439
<i>Ariyam Das and Harish Siddapura Ranganath</i>	
Author Index	441