

2015 IEEE International Congress on Big Data

(BigData Congress 2015)

**New York City, New York, USA
27 June - 2 July 2015**



**IEEE Catalog Number: CFP15SEV-POD
ISBN: 978-1-4673-7279-4**

2015 IEEE International Congress on Big Data

BigDataCongress 2015

Table of Contents

Message from the General Chairs	xvii
Message from the Program Committee	
Chairs	xviii
Organizing Committee	xix
Program Committee	xxi
External Reviewers	xxv
IEEE Computer Society Technical Committee on Services Computing	xxvi

Research Track

Research Session 1: Mining I

A Scalable Approach for Online Hierarchical Big Data Mining	1
<i>N. Denizcan Vanli, Muhammed O. Sayin, Ibrahim Delibalta, and Suleyman S. Kozat</i>	
A Parallel Distributed Weka Framework for Big Data Mining Using Spark	9
<i>Aris-Kyriakos Koliopoulos, Paraskevas Yiapanis, Firat Tekiner, Goran Nenadic, and John Keane</i>	
Geometrical and Topological Modelling: A Fast Computation of Spatial 3D TLS Data Selections	17
<i>José I. Rodrigues, Mauro Figueiredo, Ivo Silvestre, and Cristina Veiga-Pires</i>	

Research Session 2: Mining II

PaWI: Parallel Weighted Itemset Mining by Means of MapReduce	25
<i>Elena Baralis, Luca Cagliero, Paolo Garza, and Luigi Grimaudo</i>	
Distributed SPARQL over Big RDF Data: A Comparative Analysis Using Presto and MapReduce	33
<i>Mulugeta Mammo and Srividya K. Bansal</i>	

A GPU Based SVM Method with Accelerated Kernel Matrix Calculation	41
<i>Bo Yan, Yitian Ren, and Zijiang Yang</i>	

Research Session 3: Big Data and Social Network

A Clustered Approach for Fast Computation of Betweenness Centrality in Social Networks	47
<i>Paolo Suppa and Eugenio Zimeo</i>	
A Semantic Recommender for Micro-blog Users	55
<i>Stefano Faralli, Giovanni Stilo, and Paola Velardi</i>	
Incorporating Tie Strength in Robust Social Recommendation	63
<i>Youliang Zhong, Jian Yang, and Robertus Nugroho</i>	

Research Session 4: Big Data and Social Network

Can We Rank Emotions? A Brand Love Ranking System for Emotional Terms	71
<i>Andreas Kanavos, Eleanna Kafeza, and Christos Makris</i>	
Matrix Inter-joint Factorization - A New Approach for Topic Derivation in Twitter	79
<i>Robertus Nugroho, Youliang Zhong, Jian Yang, Cecile Paris, and Surya Nepal</i>	
Deriving Topics in Twitter by Exploiting Tweet Interactions	87
<i>Robertus Nugroho, Jian Yang, Youliang Zhong, Cecile Paris, and Surya Nepal</i>	

Research Session 5: Privacy

Privacy Preserving Data Analysis in Mental Health Research	95
<i>Jingquan Li and Xueying Li</i>	
Enabling Privacy Mechanisms in Apache Storm	102
<i>Christian Schaefer and P.M. Manoj</i>	
Sensitive Disclosures under Differential Privacy Guarantees	110
<i>Chao Han and Ke Wang</i>	

Research Session 6: Big Data and Learning

Red-RF: Reduced Random Forest for Big Data Using Priority Voting & Dynamic Data Reduction	118
<i>Hussein Mohsen, Hasan Kurban, Kurt Zimmer, Mark Jenne, and Mehmet M. Dalkilic</i>	
Optimal and Efficient Distributed Online Learning for Big Data	126
<i>Muhammed O. Sayin, N. Denizcan Vanli, Ibrahim Delibalta, and Suleyman S. Kozat</i>	
Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection	134
<i>Huaming Chen, Hong Zhao, Jun Shen, Rui Zhou, and Qingguo Zhou</i>	

Research Session 7: Query

CoUPE: Continuous Query Processing Engine for Evolving Graphs	142
<i>Phani Rohit Mullangi and Lakshmesh Ramaswamy</i>	
Lightweight Distributed Execution Engine for Large-Scale Spatial Join Query Processing	150
<i>Jianting Zhang, Simin You, and Le Gruenwald</i>	
Efficient and Self-Balanced ROLLUP Aggregates for Large-Scale Data Summarization	158
<i>Duy-Hung Phan, Quang-Nhat Hoang-Xuan, Matteo Dell'Amico, and Pietro Michiardi</i>	

Research Session 8: Big Data Processing

Tree Matching Using Data Shaping	166
<i>Parijat Shukla and Arun K. Somani</i>	
A Study of Data Locality in YARN	174
<i>Yehia Elshater, Patrick Martin, Dan Rope, Mike McRoberts, and Craig Statchuk</i>	
CEPSim: A Simulator for Cloud-Based Complex Event Processing	182
<i>Wilson A. Higashino, Miriam A.M. Capretz, and Luiz F. Bittencourt</i>	

Research Session 9: Big Data Quality

Big Data Pre-processing: A Quality Framework	191
<i>Ikbale Taleb, Rachida Dssouli, and Mohamed Adel Serhani</i>	
Towards Automatic Veracity Assessment of Open Source Information	199
<i>Marianela García Lozano, Ulrik Franke, Magnus Rosell, and Vladimir Vlassov</i>	
Identifying Relevant Formal Concepts through the Collapse Index	207
<i>Daniel Joseph, Nikolay Mehandjiev, Babis Theodoulidis, John Davies, and Ian Thurlow</i>	

Research Session 10: Big Data Platform/Framework

Cleaning Framework for Big Data - Object Identification and Linkage	215
<i>Hong Liu, T.K. Ashwin Kumar, and Johnson P. Thomas</i>	
Performance Analysis of Scheduling Algorithms for Dynamic Workflow Applications	222
<i>Chaochao Zhou and Saurabh Kumar Garg</i>	
Hybrid Traffic Speed Modeling and Prediction Using Real-World Data	230
<i>Rong Zhang, Yuanchao Shu, Zequ Yang, Peng Cheng, and Jiming Chen</i>	

Research Session 11: Big Data Semantics

A Big Data Modeling Methodology for Apache Cassandra	238
<i>Artem Chebotko, Andrey Kashlev, and Shiyong Lu</i>	
Replica Placement in Multi-tenant Database Environments	246
<i>Avrilia Floratou and Jignesh M. Patel</i>	
Using Semantics in Predictive Big Data Analytics	254
<i>Mustafa V. Nural, Michael E. Cotterell, and John A. Miller</i>	

Research Session 12: Analysis on Big Data Research and Platforms

Meta Data: Big Data Research Evolving across Disciplines, Players, and Topics	262
<i>Alan L. Porter, Ying Huang, Jannik Schuehle, and Jan Youtie</i>	
Big Data Open Source Platforms	268
<i>Pedro Daniel Coimbra de Almeida and Jorge Bernardino</i>	

Data Science Special Track

DS Session 1

Intelligent Big Data Analysis Architecture Based on Automatic Service Composition	276
<i>T.H.A.S. Siriweera, Incheon Paik, Banage T.G.S. Kumara, and K.R.C. Koswatta</i>	

DS Session 2

Design and Implementation of Network Transfer Protocol for Big Genomic Data	281
<i>Mohammed Aledhari and Fahad Saeed</i>	
Optimizing Query Prices for Data-as-a-Service	289
<i>Ana Cristina Oliveira, Christof Fetzer, André Martin, and Marco Spohn</i>	
Supporting Performance Isolation in Software as a Service Systems with Rich Clients	297
<i>Alp Oral and Bedir Tekinerdogan</i>	

Big Data Research in Healthcare Special Track

BDRH Session 1

Blood Pressure Management with Data Capturing in the Cloud among Hypertensive Patients: A Monitoring Platform for Hypertensive Patients	305
<i>Benjamin Yip, Hoyee W. Hirai, Yong-Hong Kuo, Helen M. Meng, Samuel Wong, and Kelvin K.F. Tsoi</i>	
Indoor Air Monitoring Platform and Personal Health Reporting System: Big Data Analytics for Public Health Research	309
<i>Kin-Fai Ho, Hoyee W. Hirai, Yong-Hong Kuo, Helen M. Meng, and Kelvin K.F. Tsoi</i>	
Embracing Big Data for Simulation Modelling of Emergency Department Processes and Activities	313
<i>Yong-Hong Kuo, Janny M.Y. Leung, Kelvin K.F. Tsoi, Helen M. Meng, and Colin A. Graham</i>	

BDRH Session 2

Risk-Adjusted Monitoring Method for Surgical Data: Methodology for Data Analytics (Work in Progress)	317
<i>Xin Lai, Liu Liu, Paul B.S. Lai, Kelvin Tsoi, Haitian Wang, Ka Chun Chong, and Benny Zee</i>	
Patient Flow Evaluation with System Dynamic Model in an Emergency Department: Data Analytics on Daily Hospital Records	320
<i>Marc Chong, Maggie Wang, Xin Lai, Benny Zee, Fung Hong, Ek Yeoh, Eliza Wong, Carrie Yam, Patsy Chau, Kelvin Tsoi, and Colin Graham</i>	
Two Screening Methods for Genetic Association Study with Application to Psoriasis Microarray Data Sets	324
<i>Maggie Haitian Wang, Kelvin Tsoi, Xin Lai, Marc Chong, Benny Zee, Tian Zheng, Shaw-Hwa Lo, and Inchi Hu</i>	

Shenzhen Satellite Track

Shenzhen Satellite Session 1

An Approach to Social Relationship Ranking on Internet-Based Social Platforms by Tempo-spatial Data Mining Using Location Prediction Technique	327
<i>Chao Ma, Yinda Wang, Haowen Liu, Hao Gui, Weiping Zhu, Xiaochuan Shi, and Xuhui Li</i>	
Optimization Approach to Depot Location in Car Sharing Systems with Big Data	335
<i>Xiaolu Zhu, Jinglin Li, Zhihan Liu, and Fangchun Yang</i>	

Hydrological Big Data Prediction Based on Similarity Search and Improved BP Neural Network	343
<i>Dingsheng Wan, Yan Xiao, Pengcheng Zhang, and Hareton Leung</i>	
An Effective Process Mining Approach against Diverse Logs Based on Case Classification	351
<i>Liqin Yang, Guosheng Kang, Weigang Cai, and Qiang Zhou</i>	

Taipei Satellite Track

Taipei Satellite Session 1

Web Service Recommendations Based on Time-Aware Bayesian Networks	359
<i>Victor W. Chu, Raymond K. Wong, Fang Chen, and Chi-Hung Chi</i>	
Crowdsourcing Service Design for Social Enterprise Insight Innovation	367
<i>Wei-Feng Tung and Guillaume Jordann</i>	
Effects of Word Assignment in LDA for News Topic Discovery	374
<i>Chuen-Min Huang and Cheng-Yi Wu</i>	
Using Big Data for Profiling Heavy Users in Top Video Apps	381
<i>Chieh-Hsin Liao, Yu-Heng Lei, Kai-Yu Liou, Jian-Shing Lin, and Hsiao-Feng Yeh</i>	
Machine Learning-Based Configuration Parameter Tuning on Hadoop System	386
<i>Chi-Ou Chen, Ye-Qi Zhuo, Chao-Chun Yeh, Che-Min Lin, and Shih-Wei Liao</i>	
Sequence-Growth: A Scalable and Effective Frequent Itemset Mining Algorithm for Big Data Based on MapReduce Framework	393
<i>Yen-Hui Liang and Shioh-Yang Wu</i>	

Application Track

Applications Session 1: Big Data and Health

Big Data Analytics Framework for System Health Monitoring	401
<i>Brian Xu and Sathish Alampalayam Kumar</i>	
Predictive Modeling for Comfortable Death Outcome Using Electronic Health Records	409
<i>Muhammad Kamran Lodhi, Rashid Ansari, Yingwei Yao, Gail M. Keena, Diana J. Wilkie, and Ashfaq A. Khokhar</i>	
H-DRIVE: A Big Health Data Analytics Platform for Evidence-Informed Decision Making	416
<i>Ashraf Abusharekh, Samuel A. Stewart, Nima Hashemian, and Syed Sibte Raza Abidi</i>	

Applications Session 2: Big Data and Network Management

Design and Realization of Cognitive Routing Resources Using Big Data Analysis in SDN	424
<i>Hongyan Cui, Yuchen Zhang, Chenhang Ma, Wei Lai, Norman C. Beaulieu, Stanislav Sobolevsky, and Yunjie Liu</i>	
Enterprise Search with Development for Network Management System	430
<i>Mingxue Wang, Robin Grindrod, Jimmy O'Meara, Mikel Zuzuarregui, Eloy Martinez, and Enda Fallon</i>	
Toa: A Web Based Network Flow Data Monitoring System at Scale	438
<i>José Ortiz-Ubarri, Humberto Ortiz-Zuazaga, Albert Maldonado, Eric Santos, and Jhensen Grullón</i>	

Applications Session 3: Distributed Processing

k-Means Performance Improvements with Centroid Calculation Heuristics Both for Serial and Parallel Environments	444
<i>Jeyhun Karimov, Murat Ozbayoglu, and Erdogan Dogdu</i>	
Graft: Arbitrary Fault-Tolerant Distributed Graph Processing	452
<i>Daniel Presser, Lau Cheuk Lung, and Miguel Correia</i>	

Applications Session 4: Social Network

A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks	460
<i>Bilal Abu Salih, Pornpit Wongthongtham, Seyed-Mehdi-Reza Beheshti, and Dengya Zhu</i>	
Efficient Algorithms for Social Network Coverage and Reach	467
<i>Deepak Puthal, Surya Nepal, Cecile Paris, Rajiv Ranjan, and Jinjun Chen</i>	
How to Choose a Recommender System: Insights and Experiences for Large-Scale User Personalization	475
<i>Rohit Parimi, Tomáš Trepka, Doina Caragea, and Cody Bennett</i>	

Applications Session 5: Social Media

Alexandria: Extensible Framework for Rapid Exploration of Social Media	483
<i>Fenno F. Heath III, Richard Hull, Elham Khabiri, Matthew Riemer, Noi Sukaviriya, and Roman Vaculín</i>	
A Latent Semantic Pattern Recognition Strategy for an Untrivial Targeted Advertising	491
<i>Roberto Saia, Ludovico Boratto, and Salvatore Carta</i>	

Applications Session 6: Image Processing

Proof of Storage for Video Deduplication in the Cloud	499
<i>Fatema Rashid, Ali Miri, and Isaac Woungang</i>	
Hadoop Image Processing Framework	506
<i>Sridhar Vemula and Christopher Crick</i>	
A Scalable Complex Pattern Mining Framework for Global Settlement Mapping	514
<i>Ranga Raju Vatsavai</i>	

Applications Session 7: Big Data Application

Study on Corporate Governance of Stock Market in Korea: Network Analysis with Relationship of Major Shareholders	522
<i>Hyejung Moon, Jangho Park, and Sungkyung Kim</i>	
Application-Specific Evaluation of No SQL Databases	526
<i>John Klein, Ian Gorton, Neil Ernst, Patrick Donohoe, Kim Pham, and Chrisjan Matser</i>	
Optigrow: People Analytics for Job Transfers	535
<i>Dennis Wei, Kush R. Varshney, and Marcy Wagman</i>	

Applications Session 8: Optimization

Energy Signature Analysis: Knowledge at Your Fingertips	543
<i>Andrea Acquaviva, Daniele Apiletti, Antonio Attanasio, Elena Baralis, Lorenzo Bottaccioli, Federico Boni Castagnetti, Tania Cerquitelli, Silvia Chiusano, Enrico Macii, Dario Martellacci, and Edoardo Patti</i>	
Ctracer: Uncover C&C in Advanced Persistent Threats Based on Scalable Framework for Enterprise Log Data	551
<i>Kai-Fong Hong, Chien-Chih Chen, Yu-Ting Chiu, and Kuo-Sen Chou</i>	
A Community-Based Cloud Computing Caching Service	559
<i>Unekwu Idachaba and Frank Wang</i>	

Applications Session 9: Evaluation

Unsupervised Event Detection with Infinite Poisson Mixture Model	567
<i>Vinod Hegde, Milovan Krnjajic, and Alexei Pozdnoukhov</i>	
Reconstructability-Aware Filtering and Forwarding of Time Series Data in Internet-of-Things Architectures	576
<i>Apostolos Papageorgiou, Bin Cheng, and Ernő Kovacs</i>	
No SQL in Practice: A Write-Heavy Enterprise Application	584
<i>João Ricardo Lourenço, Veronika Abramova, Bruno Cabral, Jorge Bernardino, Paulo Carreiro, and Marco Vieira</i>	

Applications Session 10: Big Data Framework

Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander	592
<i>Bin Cheng, Salvatore Longo, Flavio Cirillo, Martin Bauer, and Ernoe Kovacs</i>	
Scaling of City Attractiveness for Foreign Visitors through Big Data of Human Economical and Social Media Activity	600
<i>Stanislav Sobolevsky, Iva Bojic, Alexander Belyi, Izabela Sitko, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti</i>	
Automation of the Validation, Anonymization, and Augmentation of Big Data from a Multi-year Driving Study	608
<i>Bruce Wallace, Rafik Goubran, Frank Knoefel, Shawn Marshall, Michelle Porter, Madelaine Harlow, and Akshay Puli</i>	

Applications Session 11: Big Data Use Cases

Big Data Driven Cyber Analytic System	615
<i>Vladimir Hahanov, Wajeb Gharibi, Eugenia Litvinova, and Svetlana Chumachenko</i>	
Nimbus: Tuning Filters Service on Tweet Streams	623
<i>Chien-An Lai, Jim Donahue, Aibek Musaev, and Calton Pu</i>	

Short Paper Track

Session 1

Computation and Search over Encrypted XML Documents	631
<i>Hoi Ting Poon and Ali Miri</i>	
Phishing URL Detection Using URL Ranking	635
<i>Mohammed Nazim Feroz and Susan Mengel</i>	
A Real-Time Decision Support Tool for Disaster Response: A Mathematical Programming Approach	639
<i>Yong-Hong Kuo, Janny M.Y. Leung, Helen M. Meng, and Kelvin K.F. Tsoi</i>	

Session 2

The Pig Mix Benchmark on Pig, MapReduce, and HPCC Systems	643
<i>Keren Ouaknine, Michael Carey, and Scott Kirkpatrick</i>	
Scalable Query Optimization for Efficient Data Processing Using MapReduce	649
<i>Yi Shan and Yi Chen</i>	

Weighted Finite Automata Based Image Compression on Hadoop MapReduce Framework	653
<i>U.S.N. Raju, Irlanki Sandeep, Nattam Sai Karthik, Rayapudi Siva Praveen, and Mayank Singh Sachan</i>	
Developing a Real-Time Data Analytics Framework Using Hadoop	657
<i>Sangwhan Cha and Monica Wachowicz</i>	
Content Based Image Retrieval on Hadoop Framework	661
<i>U.S.N. Raju, Shibin George, V. Sairam Praneeth, Ranjeet Deo, and Priyanka Jain</i>	

Session 3

High-Order Tensor Decomposition for Large-Scale Data Analysis	665
<i>Longzhuang Li and Douglas Boulware</i>	
Optimal Feature Extraction and Classification of Tensors via Matrix Product State Decomposition	669
<i>Johann A. Bengua, Ho N. Phien, and Hoang D. Tuan</i>	
A Workflow Model for Adaptive Analytics on Big Data	673
<i>Verena Kantere and Maxim Filatov</i>	
Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty	677
<i>Muhammad Raza Khan, Joshua Manoj, Anikate Singh, and Joshua Blumenstock</i>	
Analytics of Industrial Operational Data Inspired by Natural Language Processing	681
<i>Mariusz Kamola</i>	
Online Nonlinear Classification for High-Dimensional Data	685
<i>N. Denizcan Vanli, Huseyin Ozkan, Ibrahim Delibalta, and Suleyman S. Kozat</i>	

Session 4

Cloud Tree: A Library to Extend Cloud Services for Trees	689
<i>Yun Tian, Bojian Xu, Yanqing Ji, and Jesse Scholer</i>	
Matrix-Based XML Stream Processing Using a GPU	694
<i>Soo-Hyung Kim, Yoon-Joon Lee, and John Jaehwan Lee</i>	
A Flexible Data-Driven Approach for Execution Trace Filtering	698
<i>Kadjo Kouame, Naser Ezzati-Jivan, and Michel R. Dagenais</i>	
Supporting Data Driven Access through Automatic Keyword Extraction and Summarization	704
<i>Weijia Xu, Wei Luo, Nicholas Woodward, and Yan Zhang</i>	
Road Traffic Analytic Query Processing Based on a Timeline Modeling	708
<i>Ardi Imawan, Titus Irma Damaiyanti, and Joonho Kwon</i>	

Approximate Queries on Big Heterogeneous Data	712
<i>Verena Kantere</i>	

Session 5

Queriosity: Automated Data Exploration	716
<i>Abdul Wasay, Manos Athanassoulis, and Stratos Idreos</i>	
Write Optimization Using Asynchronous Update on Out-of-Core Column-Store Databases in Map-Reduce	720
<i>Feng Yu, Eric S. Jones, and Wen-Chi Hou</i>	
Internet Archives as a Tool for Research: Decay in Large Scale Archival Records	724
<i>Hai Nguyen and Matthew S. Weber</i>	
Social Network Analysis of Developers' and Users' Mailing Lists of Some Free Open Source Software	728
<i>Armel Jacques Nzekon Nzeko'o, Matthieu Latapy, and Maurice Tchuente</i>	
How Has Twitter Changed the Event Discussion Scenario? A Spatio-temporal Diffusion Analysis	733
<i>Purva Pruthi, Anu Yadav, Farheen Abbasi, and Durga Toshniwal</i>	

Session 6

A Data Capturing Platform in the Cloud for Behavioral Analysis among Smokers: An Application Platform for Public Health Research	737
<i>Kelvin K.F. Tsoi, Yong-Hong Kuo, and Helen M. Meng</i>	
Discovering Environmental Impacts on Public Health Using Heterogeneous Big Sensory Data	741
<i>Minh-Son Dao and Koji Zettsu</i>	
Detecting Corporate Social Media Crises on Facebook Using Social Set Analysis	745
<i>Raghava Rao Mukkamala, Jannie Iskou Sørensen, Abid Hussain, and Ravi Vatrappu</i>	
Data Quality and Energy Management Tradeoffs in Sensor Service Clouds	749
<i>Victor Lawson and Lakshmish Ramaswamy</i>	
Two Steps Genetic Programming for Big Data - Perspective of Distributed and High-Dimensional Data	753
<i>Jih-Jeng Huang</i>	

Visionary Track

Visionary Session

Big Data - Security and Privacy	757
<i>Elisa Bertino</i>	
MCD: Mutual Clustering across Multiple Social Networks	762
<i>Philip S. Yu and Jiawei Zhang</i>	
Database Security: Past, Present, and Future	772
<i>Bhavani Thuraisingham</i>	
Big SaaS: The Next Step Beyond Big Data	775
<i>Hong Zhu, Ian Bayley, M. Younas, David Lightfoot, Basel Yousef, and Dongmei Liu</i>	
Research Directions for Big Data Graph Analytics	785
<i>John A. Miller, Lakshmish Ramaswamy, Krys J. Kochut, and Arash Fard</i>	
Author Index	795