

2016 IEEE International Congress on Big Data (BigData Congress 2016)

**San Francisco, California, USA
27 June – 2 July 2016**



**IEEE Catalog Number: CFP16SEV-POD
ISBN: 978-1-5090-2623-4**

**Copyright © 2016 by the Institute of Electrical and Electronics Engineers, Inc
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

******This publication is a representation of what appears in the IEEE Digital Libraries. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP16SEV-POD
ISBN (Print-On-Demand):	978-1-5090-2623-4
ISBN (Online):	978-1-5090-2622-7
ISSN:	2379-7703

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2016 IEEE International Congress on Big Data

BigDataCongress 2016

Table of Contents

Message from the General Chairs.....	xii
Message from the Program Committee	
Chairs.....	xiv
Organizing Committee.....	xv
Program Committee.....	xviii
IEEE Computer Society Technical Committee on Services Computing.....	xxv

BigData Congress Research Track

Session 1: Best Paper Session 1

Boosting Vertex-Cut Partitioning for Streaming Graphs	1
<i>Hooman Peiro Sajjad, Amir H. Payberah, Fatemeh Rahimian, Vladimir Vlassov, and Seif Haridi</i>	
Optimizing Hadoop Framework for Solid State Drives	9
<i>Jaeki Hong, Liang Li, Chihye Han, Bingxu Jin, Qichao Yang, and Zilong Yang</i>	
Continual and Cost-Effective Partitioning of Dynamic Graphs for Optimizing Big Graph Processing Systems	18
<i>Amirreza Abdolrashidi and Lakshmish Ramaswamy</i>	

Session 2: Best Paper and Big Data for Enterprise

Evaluation and Analysis of In-Memory Key-Value Systems	26
<i>Wenqi Cao, Semih Sahin, Ling Liu, and Xianqiang Bao</i>	
Improving the Visualization of WordNet Large Lexical Database through Semantic Tag Clouds	34
<i>Enrico Giacinto Caldarola and Antonio M. Rinaldi</i>	
Clustering Geo-tagged Tweets for Advanced Big Data Analytics	42
<i>Gloria Bordogna, Luca Frigerio, Alfredo Cuzzocrea, and Giuseppe Psaila</i>	

Session 3: Big Data Models and Algorithms

A NoSQL Data Model for Scalable Big Data Workflow Execution	52
<i>Aravind Mohan, Mahdi Ebrahimi, Shiyong Lu, and Alexander Kotov</i>	
Model Transformation and Data Migration from Relational Database to MongoDB	60
<i>Tianyu Jia, Xiaomeng Zhao, Zheng Wang, Dahan Gong, and Guiguang Ding</i>	
Predictive Modeling in a Big Data Distributed Setting: A Scalable Bias Correction Approach	68
<i>Gianluca Bontempi and Yann-Aël Le Borgne</i>	

Session 4: Big Data Processing

Distributed Incremental Graph Analysis	75
<i>Upa Gupta and Leonidas Fegaras</i>	
Enhanced State History Tree (eSHT): A Stateful Data Structure for Analysis of Highly Parallel System Traces	83
<i>Loïc Prieur-Drevon, Raphaël Beamonte, Naser Ezzati-Jivan, and Michel R. Dagenais</i>	
A Case Study of Optimizing Big Data Analytical Stacks Using Structured Data Shuffling	91
<i>Dixin Tang, Taoying Liu, Rubao Lee, Hong Liu, and Wei Li</i>	

Session 5: Big Data Analytics and Metrics

Geelytics: Enabling On-Demand Edge Analytics over Scoped Data Sources	101
<i>Bin Cheng, Apostolos Papageorgiou, and Martin Bauer</i>	
Analytics Toolkit for Business Big Data	109
<i>Fan Liang and Weichang Du</i>	
QDrill: Query-Based Distributed Consumable Analytics for Big Data	117
<i>Shadi Khalifa, Patrick Martin, Dan Rope, Mike McRoberts, and Craig Statchuk</i>	

Session 6: Big Data Applications

Marketing Campaigns in Twitter Using a Pattern Based Diffusion Policy	125
<i>Eleanna Kafeza, Christos Makris, and Pantelis Vikatos</i>	
TwigStack-MR: An Approach to Distributed XML Twig Query Using MapReduce	133
<i>Hongjie Fan, Han Yang, Zhiyi Ma, and Junfei Liu</i>	
Infra: SLO Aware Elastic Auto-scaling in the Cloud for Cost Reduction	141
<i>Subhajit Sidhanta and Supratik Mukhopadhyay</i>	

Session 7: Big Data Management

GOMA: Supporting Big Data Analytics with a Goal-Oriented Approach	149
<i>Sam Supakkul, Liping Zhao, and Lawrence Chung</i>	
Energy Consumption Prediction with Big Data: Balancing Prediction Accuracy and Computational Resources	157
<i>Katarina Grolinger, Miriam A.M. Capretz, and Luke Seewald</i>	
From Big Data to Great Services	165
<i>Jianwei Yin, Yan Tang, Wei Lo, and Zhaohui Wu</i>	

Session 8: Big Data Privacy

A Fast and Efficient Entity Resolution Approach for Preserving Privacy in Mobile Data	173
<i>Ioannis Boutsis and Vana Kalogeraki</i>	
SD-HDFS: Secure Deletion in Hadoop Distributed File System	181
<i>Bikash Agrawal, Raymond Hansen, Chunming Rong, and Tomasz Wiktorski</i>	
Privacy Preserving Predictive Analytics with Smart Meters	190
<i>B. Habtemariam, A. Miranskyy, A. Miri, S. Samet, and M. Davison</i>	

Session 9: Big Data Search and Mining

Catching Social Butterflies: Identifying Influential Users of an Event-Based Social Networking Service	198
<i>Jonathan Popa, Kusha Nezafati, Yulia R. Gel, John Zweck, and Georgiy Bobashev</i>	
Towards an Efficient Top-K Trajectory Similarity Query Processing Algorithm for Big Trajectory Data on GPGPUs	206
<i>Eleazar Leal, Le Gruenwald, Jianting Zhang, and Simin You</i>	
DeepSky: Identifying Absorption Bumps via Deep Learning	214
<i>Xiaoyong Yuan, Min Li, Sudeep Gaddam, Xiaolin Li, Yinan Zhao, Jingzhe Ma, and Jian Ge</i>	

BigData Congress Application Track

Session 1: Big Data Applications

Green Cabs vs. Uber in New York City	222
<i>Lasse Korsholm Poulsen, Daan Dekkers, Nicolaas Wagenaar, Wesley Snijders, Ben Lewinsky, Raghava Rao Mukkamala, and Ravi Vatrapu</i>	
Driver Unique Acceleration Behaviours and Stability over Two Years	230
<i>Bruce Wallace, Rafik Goubran, Frank Knoefel, Shawn Marshall, Michelle Porter, and Andrew Smith</i>	

Session 2: Text Processing

Cloud-Based Core Text Processing Services for Sentiment Analysis	236
<i>Huan Chen, Xin-Nan Li, Liang-Jie Zhang, Yi-Xuan Huang, and Xiao-Sheng Cai</i>	
Design and Implementation of a Multidimensional Data Retrieval Sorting Optimization Model	244
<i>Danfeng Yan, Liying Zhang, and Xuan Zhao</i>	
An Improved Single-Pass Algorithm for Chinese Microblog Topic Detection and Tracking	251
<i>Danfeng Yan, Enzheng Hua, and Bo Hu</i>	

Session 3: Big Data Streaming

Edge-Computing-Aware Deployment of Stream Processing Tasks Based on Topology-External Information: Model, Algorithms, and a Storm-Based Prototype	259
<i>Apostolos Papageorgiou, Ehsan Poormohammady, and Bin Cheng</i>	
Effective Multi-stream Joining in Apache Samza Framework	267
<i>Zhenyun Zhuang, Tao Feng, Yi Pan, Haricharan Ramachandra, and Badri Sridharan</i>	

Session 4: Evaluation

A Scalable Probabilistic Change Detection Algorithm for Very High Resolution (VHR) Satellite Imagery	275
<i>Seokyoung Hong and Ranga Raju Vatsavai</i>	
Evaluation of Data Mining Tools for Telecommunication Monitoring Data Using Design of Experiment	283
<i>Samneet Singh, Yan Liu, Wayne Ding, and Zheng Li</i>	
An Empirical Investigation of Mobile Network Traffic Data for Resource Management	291
<i>Man Si, Chung-Horng Lung, Samuel Ajila, and Wayne Ding</i>	

Session 5: Big Data Privacy, Security, and Economics

Genetic Information Privacy in the Age of Data-Driven Medicine	299
<i>Jingquan Li</i>	
A Method to Price Your Information Asset in the Information Market	307
<i>Divya Rao and Ng Wee Keong</i>	

BigData Congress Short Paper Track

Session 1: Big Data Processing

Workflow Transformation for Real-Time Big Data Processing	315
<i>Yuji Ishizuka, Wuhui Chen, and Incheon Paik</i>	
An OWL Ontology Representation for Machine-Learned Functions Using Linked Data	319
<i>Jingyuan Xu, Hao Wang, and Henry Trimbach</i>	
Quantitative Analysis of Scalable NoSQL Databases	323
<i>Surya Narayanan Swaminathan and Ramez Elmasri</i>	

Session 2: Big Data Management

Asset-centric Security-Aware Service Selection	327
<i>Giannis Tziakouris, Marios Zinonos, Tom Chothia, and Rami Bahsoon</i>	
Semi-clustering That Scales: An Empirical Evaluation of GraphX	333
<i>Jakob Smedegaard Andersen and Olaf Zukunft</i>	
Monetizing the User's Information Asset in Internet Information Market	337
<i>Divya Rao and Wee Keong Ng</i>	

Session 3: Big Data Systems and Architecture

Privacy-Aware Big Data Warehouse Architecture	341
<i>Karthik Navuluri, Ravi Mukkamala, and Aftab Ahmad</i>	
Software Metrics for Green Parallel Computing of Big Data Systems	345
<i>Havva Gülay Gürbüz and Bedir Tekinerdogan</i>	

Session 4: Big Data Systems and Mining

Distributed Top-k Keyword Search over Very Large Databases with MapReduce	349
<i>Ziqiang Yu, Xiaohui Yu, Yuehui Chen, and Kun Ma</i>	
Towards Automatic Memory Tuning for In-Memory Big Data Analytics in Clusters	353
<i>Aris-Kyriakos Koliopoulos, Paraskevas Yiapanis, Firat Tekiner, Goran Nenadic, and John Keane</i>	
Analyzing Future Nodes in a Knowledge Network	357
<i>Sukhwan Jung, Tuan Manh Lai, and Aviv Segev</i>	

BigData Congress Special Track

Session 1: Big Data Applications

Do Drivers' Behaviors Reflect Their Past Driving Histories? - Large Scale Examination of Vehicle Recorder Data	361
<i>Daisaku Yokoyama and Masashi Toyoda</i>	
Urban Area Vehicle Number Estimation Based on RTMS Data	369
<i>Yue Hu, Yuanchao Shu, Peng Cheng, and Jiming Chen</i>	
Towards Data Analytics of Pathogen-Host Protein-Protein Interaction: A Survey	377
<i>Huaming Chen, Jun Shen, Lei Wang, and Jiangning Song</i>	

Session 2: Big Data Management

Open Source Big Data Analytics Frameworks Written in Scala	389
<i>John A. Miller, Casey Bowman, Vishnu Gowda Harish, and Shannon Quinn</i>	
Big Datasets for Research: A Survey on Flagship Conferences	394
<i>Yi Wei, Shijun Liu, Jiao Sun, Lizhen Cui, Li Pan, and Lei Wu</i>	
Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM	402
<i>Ali Yadavar Nikravesh, Samuel A. Ajila, Chung-Horng Lung, and Wayne Ding</i>	

Session 3: Evaluation and Operations

JVM Configuration Management and Its Performance Impact for Big Data Applications	410
<i>Semih Sahin, Wenqi Cao, Qi Zhang, and Ling Liu</i>	
An Hybrid Approach to Quality Evaluation across Big Data Value Chain	418
<i>Mohamed Adel Serhani, Hadeel T. El Kassabi, Ikbal Taleb, and Alramzana Nujum</i>	

BigData Congress Services Special Track

Session 1: Service Applications

GS1 Connected Car Using EPCIS-ONS System	426
<i>Bongjin Sohn, Sungpil Woo, Jiyong Han, Hyeemun Cho, Jaewook Byun, and Daeyoung Kim</i>	
Key Update as a Service (KAAS): An Agent-Based Modeling for Cloud-Based Access Control	430
<i>Somchart Fugkeaw and Hiroyuki Sato</i>	

Session 2: Service Modeling

Modeling the Location Selection of Mirror Servers in Content Delivery Networks	438
<i>Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, and Oliver Rose</i>	
Cloud Computing Intelligent Data-Driven Model: Connecting the Dots to Combat Global Terrorism	446
<i>Gokop L. Goteng and Xueyu Tao</i>	
Research of Clouding Based Pulse Monitoring and Data Analysis Framework for Efficient Physical Training	454
<i>Hong Liang Yuan, Jun Wang, Jun Liu, and Shi Liang Li</i>	

Session 3: QoS Management

Complex Quality of Service Lifecycle Assessment Methodology	462
<i>R. William Maule</i>	
Don't Fire Me, a Kernel Autoregressive Hybrid Model for Optimal Layoff Plan	470
<i>Zhilong Luo, Ying Li, Ruisheng Fu, and Jianwei Yin</i>	
Can We Use Programmer's Knowledge? Fixing Parameter Configuration Errors in Hadoop through Analyzing Q&A Sites	478
<i>Jia Tong, Li Ying, Tang Hongyan, and Wu Zhonghai</i>	

Session 4: Services Clustering and Discovery

Identification as a Service: Large-Scale Cloud Service Discovery over the World Wide Web	485
<i>Abdullah Alfazi, Quan Z. Sheng, Wei Emma Zhang, Lina Yao, and Talal H. Noor</i>	
Software Evolution Information Driven Service-Oriented Software Clustering	493
<i>Linhui Zhong, Jing He, Nengwei Zhang, Peng Zhang, and Jing Xia</i>	
Author Index	501