2014 IEEE International Congress on Big Data

(BigData Congress 2014)

Anchorage, Alaska, USA
27 June – 2 July 2014
Session 1

BigData Research Session 1 - BigData Analytics Techniques

GraphLens: Mining Enterprise Storage Workloads Using Graph Analytics .............................................1
  Yang Zhou, Sangeetha Seshadri, Lawrence Chiu, and Ling Liu

FSM-H: Frequent Subgraph Mining Algorithm in Hadoop ........................................................................9
  Mansurul A. Bhuiyan and Mohammad Al Hasan

Rectangle Counting in Large Bipartite Graphs .....................................................................................17
  Jia Wang, Ada Wai-Chee Fu, and James Cheng

BigData Research Session 2 - MapReduce Model

A Parallel Spatial Co-location Mining Algorithm Based on MapReduce .............................................25
  Jin Soung Yoo, Douglas Boulware, and David Kimmey

Energy-Aware Scheduling of MapReduce Jobs .....................................................................................32
  Lena Mashayekhy, Mahyar Movahed Nejad, Daniel Grosu, Dajun Lu,
  and Weisong Shi

Vigiles: Fine-Grained Access Control for MapReduce Systems .........................................................40
  Huseyin Ulusoy, Murat Kantarcioglu, Erman Pattuk, and Kevin Hamlen
BigData Research Session 3 - BigData Security

Denial-of-Service Threat to Hadoop/YARN Clusters with Multi-tenancy .................................................................48
  Jingwei Huang, David M. Nicol, and Roy H. Campbell

A Big Data Architecture for Large Scale Security Monitoring .....................................................................................56
  Samuel Marchal, Xiuyan Jiang, Radu State, and Thomas Engel

Contextual Anomaly Detection in Big Sensor Data ........................................................................................................64
  Michael A. Hayes and Miriam A.M. Capretz

BigData Research Session 4 - BigData Performance Issues

High-Performance Spatial Query Processing on Big Taxi Trip Data Using GPGPUs ........................................................72
  Jianting Zhang, Simin You, and Le Gruenwald

A Compatible LZMA ORC-Based Optimization for High Performance Big Data Load .......................................................80
  Liping Zhang, Qi Chen, and Kai Miao

BigData Research Session 5 - BigData Analytics

Storing a Collection of Differentially Compressed Files Recursively ........................................................................88
  Angelos Molfetas, Anthony Wirth, and Justin Zobel

XDB - A Novel Database Architecture for Data Analytics as a Service ...........................................................................96
  Carsten Binnig, Abdallah Salama, Erfan Zamanian, Harald Kommayer, Sven Listing, and Alexander C. Mueller

DeltaDB: A Scalable Database Design for Time-Varying Schema-Free Data ..............................................................104
  Peter Ivie and Douglas Thain

BigData Research Session 6 - MapReduce Framework

Practising Scalable Graph Similarity Joins in MapReduce ..........................................................................................112
  Yifan Chen, Xiang Zhao, Bin Ge, Chuan Xiao, and Chi-Hung Chi

Configuring a MapReduce Framework for Performance-Heterogeneous Clusters ......................................................120
  Jessica Hartog, Renan Delvalle, Madhusudhan Govindaraju, and Michael J. Lewis

Marimba: A Framework for Making MapReduce Jobs Incremental ........................................................................128
  Johannes Schildgen, Thomas Jörg, Manuel Hoffmann, and Stefan Deßloch
BigData Research Session 7 - BigData Services
Money on the Move: Big Data of Bank Card Transactions as the New Proxy for Human Mobility Patterns and Regional Delineation. The Case of Residents and Foreign Visitors in Spain .................................................................................................................. 136
  Stanislav Sobolevsky, Izabela Sitko, Remi Tachet Des Combes, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti
Characterizing and Modeling Package Dynamics in Express Shipping Service Network ..................................................................................................................................................... 144
  Xu Tan, Yuanchao Shu, Xie Lu, Peng Cheng, and Jiming Chen
Dmodel: Online Taxicab Demand Model from Big Sensor Data in a Roving Sensor Network ........................................................................................................................................................................ 152
  Desheng Zhang, Tian He, Shan Lin, Sirajum Munir, and John A. Stankovic

BigData Research Session 8 - Distributed BigData Services
Distributed Maximal Clique Computation .......................................................................................................................................................................................... 160
  Yanyan Xu, James Cheng, Ada Wai-Chee Fu, and Yingyi Bu
A Processing Pipeline for Cassandra Datasets Based on Hadoop Streaming .............................................................................................................. 168
  E. Dede, B. Sendir, P. Kuzlu, J. Weachock, M. Govindaraju, and L. Ramakrishnan
AuDy: Automatic Dynamic Least-Weight Balancing for Stream Workloads Scalability .................................................................................................................................................................................................................. 176
  Pedro Martins, Maryam Abbasi, and Pedro Furtado

BigData Research Session 9 - BigData Analytics Network
SiLK: A Tool Suite for Unsampled Network Flow Analysis at Scale ............................................................................................................................. 184
  Mark Thomas, Leigh Metcalf, Jonathan Spring, Paul Krystosek, and Katherine Prevost
Using Inter-file Similarity to Improve Intra-file Compression ........................................................................................................................... 192
  Angelos Molfetas, Anthony Wirth, and Justin Zobel

BigData Research Session 10 - BigData Management Model
A Cloud-Computing Local Histogram Construction Algorithm for Big Image Data .................................................................................................................................................................................................................................. 200
  Chung-Chih Cheng, Fan-Chieh Cheng, Po-Hsiung Lin, and Shih-Chia Huang
Scalable State Management for Scientific Applications in the Cloud ...................................................................................................................... 204
  Tonglin Li, Ioan Raicu, and Lavanya Ramakrishnan
T-PICE: Twitter Personality Based Influential Communities Extraction System .................................................................................................. 212
  Eleanna Kafeza, Andreas Kanavos, Christos Makris, and Pantelis Vikatos
A Holistic Framework for Big Scientific Data Management .................................................................220
Verena Kantere

**BigData Research Session 11 - Mexico City Satellite Session**

Data Base Analysis Using a Compact Data Set ..........................................................................................227
Angel Fernando Kuri-Morales

The Babel File System ...............................................................................................................................234
Moisés Quezada Naquid, Ricardo Marcelín Jiménez, and Jose Luis González Compeán

**BigData Research Session 12 - Coimbra Satellite Session**

Using Multimedia Ontologies for Automatic Image Annotation and Classification ..................................242
Antonio M. Rinaldi

Real-Time Integration of Building Energy Data .......................................................................................250
Diogo Anjos, Paulo Carreira, and Alexandre P. Francisco

**BigData Research Session 13 - Taipei Satellite Session**

Provenance as a Service: A Data-centric Approach for Real-Time Monitoring .......................................258
Rafat Hammad and Ching-Seh Wu

Big Data Collection Gateway for Vision-Based Smart Meter Reading Network ......................................266
Ching-Han Chen, Ching-Yi Chen, Chih-Hsien Hsia, and Guan-Xin Wu

Triangulation-Based Singer Identification for Duet Music Data Indexing ..............................................270
Wei-Ho Tsai and Cin-Hao Ma

Speech and Singing Discrimination for Audio Data Indexing ................................................................276
Wei-Ho Tsai and Cin-Hao Ma

Temporal Event Tracing on Big Healthcare Data Analytics ......................................................................281
Chin-Ho Lin, Liang-Cheng Huang, Seng-Cho T. Chou, Chih-Ho Liu, Han-Fang Cheng, and I-Jen Chiang

The Application of Cartesian-Join of Bloom Filters to Supporting Membership Query of Multidimensional Data .................................................................................................................288
Zhu Wang, Tiejian Luo, Guandong Xu, and Xiang Wang

Federated MapReduce to Transparently Run Applications on Multicluster Environment ................................296
Chun-Yu Wang, Tzu-Li Tai, Shu Jui-Shing, Chang Jyh-Biau, and Shieh Ce-Kuen

How to Send a Self-Destructing Email: A Method of Self-Destructing Email System ................................304
Xiao Fu, Zhijian Wang, Hao Wu, Jia-qi Yang, and Zi-zhao Wang
Association Rule Mining of Personal Hobbies in Social Networks ........................................310
    Xiaoqing Yu, Huanhuan Liu, Jianhua Shi, Jenq-Neng Hwang, Wanggen Wan,
    and Jing Lu

**BigData Research Session 14 - Data Processing**

Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data ..........................................................315
    Carson Kai-Sang Leung, Richard Kyle MacKinnon, and Fan Jiang

Lethe: Cluster-Based Indexing for Secure Multi-user Search ..........................................................323
    Eirini C. Micheli, Giorgos Margaritis, and Stergios V. Anastasiadis

Adaptive Indexing for Distributed Array Processing ..........................................................331
    Yifeng Geng, Xiaomeng Huang, and Guangwen Yang

**BigData Research Session 15 - Shenzhen Satellite Session**

Hydrological Time Series Anomaly Mining Based on Symbolization and Distance Measure ..........................................................339
    Dingsheng Wan, Yan Xiao, Pengcheng Zhang, Jun Feng, Yuelong Zhu,
    and Qian Liu

Enterprise Relationship Network: Build Foundation for Social Business ..........................................................347
    Liqiang Wang, Shijun Liu, Li Pan, Lei Wu, and Xiangxu Meng

Towards Big Data Bayesian Network Learning - An Ensemble Learning Based Approach ..........................................................355
    Yan Tang, Yu Wang, Kendra M.L. Cooper, and Ling Li

A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec ..........................................................358
    Bai Xue, Chen Fu, and Zhan Shaobin

A Study on Trend Prediction in Sina Weibo Community ..........................................................364
    Chen Fu, Zhan Shaobin, and Shi Guangjin

MapCheckReduce: An Improved MapReduce Computing Model for Imprecise Applications ..........................................................366
    Changjian Wang, Yuxing Peng, Mingxing Tang, Dongsheng Li, Shanshan Li,
    and Pengfei You

Three-Level Views of the Web Service Network: An Empirical Study Based on ProgrammableWeb ..........................................................374
    Saixia Lyu, Jianxun Liu, Mingdong Tang, Guosheng Kang, Buqing Cao,
    and Yucong Duan

Predicting Wikipedia Editor's Editing Interest Based on Factor Graph Model ..........................................................382
    Haisu Zhang, Sheng Zhang, Zhaolin Wu, Liwei Huang, and Yutao Ma
On Retrieving Moving Objects Gathering Patterns from Trajectory Data
via Spatio-temporal Graph

Junming Zhang, Jinglin Li, Shangguang Wang, Zhihan Liu, Quan Yuan,
and Fangchun Yang

BigData Industry and Application Session 1 - Distributed Data Model

BigData Oriented Open Scalable Relational Data Model

Zhiyun Zheng, Zhimeng Du, Lun Li, and Yike Guo

ADAM - A Database and Information Retrieval System for Big Multimedia Collections

Ivan Giangreco, Ihab Al Kabary, and Heiko Schuldt

Building a Massive Stream Computing Platform for Flexible Applications

Tianjian Chen, Zhengrui Man, Hao Li, Xin Sun, Raymond K. Wong, and Zhiwei Yu

BigData Industry and Application Session 2 - Cloud-Based BigData Management

OpenStack Café: A Novel Time-Based User-centric Resource Management Framework in the Cloud

Alan Yu Shyang Tan, Ryan Kok Leong Ko, and Grace P.Y. Ng

Trend Analysis of Time Series Data Using Data Mining Techniques

Arpit Baheti and Durga Toshniwal

Parallel POD Compression of Time-Varying Big Datasets Using m-Swap on the K Computer

Chongke Bi, Kenji Ono, and Lu Yang

BigData Industry and Application Session 3 - BigData Management System

A Key-Value Based Application Platform for Enterprise Big Data

Bo Hu, Yutao Ma, Liang-Jie Zhang, Jiake Shi, and Jiayan Zhong

Does Social Media Big Data Make the World Smaller? An Exploratory Analysis of Keyword-Hashtag Networks

Ahmed Abdeen Hamed and Xindong Wu

Data Visualization in Educational Datasets Using a Rule-Based Inference System

Aniruddha Desai, Muaz Mian, David Hazel, Ankur Teredesai, and Gregory Benner
BigData Industry and Application Session 4 - BigData Analytics Method

A Rule-Based Temporal Analysis Method for Online Health Analytics and Its Application for Real-Time Detection of Neonatal Spells .................................................................470

Anirudh Thommandram, J. Mikael Eklund, Carolyn McGregor,
James Edward Pugh, and Andrew G. James

A Discriminant Framework for Detecting Similar Scientific Research Projects Based on Big Data Mining ..................................................................................................................478

Shanqing Li, Lirong Song, and Hui Zhao

BigData Industry and Application Session 5 - BigData Analytics Algorithm

Distributed Implementation of Latent Rating Pattern Sharing Based Cross-domain Recommender System Approach .................................................................................................482

Anil Kumar, Vikas Kapur, Apangshu Saha, Rajeev Kumar Gupta, Arun Singh,
Santanu Chaudhury, and Sumeet Agarwal

Auto-configuration System and Algorithms for Big Data-Enabled Internet-of-Things Platforms .................................................................................................................................490

Apostolos Papageorgiou, Manuel Zahn, and Ernö Kovacs

DualIso: An Algorithm for Subgraph Pattern Matching on Very Large Labeled Graphs .................................................................................................................................498

Matthew Saltz, Ayushi Jain, Abhishek Kothari, Arash Fard, John A. Miller,
and Lakshmish Ramaswamy

BigData Industry and Application Session 6 - BigData Mining

An Autonomic Mediation Framework for Complex Physical Environments .................................................................506

Philippe Lalanda and Catherine Hamon

FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing .................................................................................................................................514

Leila Ismail, Mohammad M. Masud, and Latifur Khan

Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration .................................................................................................................................522

Srividya K. Bansal

BigData Industry and Application Session 7 - NoSQL

Big Data Analytics for Predictive Manufacturing Control - A Case Study from Process Industry .................................................................................................................................530

Julian Krumeich, Sven Jacobi, Dirk Werth, and Peter Loos

BigObject Store: In-Place Computing for Interactive Analytics .................................................................................................................................538

Yi-Cheng Huang, Wenwey Hseush, Yu-Chun Lai, and Michael Fong
Scalable XPath Evaluation on Large-Scale Continuously Evolving XML Repositories
Phani Rohit Mullangi, Gowtham Penematsa, and Lakshmish Ramaswamy

**BigData Industry and Application Session 8 - Big Spatial Data**
Bridging the Gap between Spatial Data Sources and Mashup Applications
Wei Zhou, Chi-Hung Chi, Can Wang, Raymond Wong, and Chen Ding
A Combinatorial Approach to Multidimensional Scaling
Jorge Alencar, Tibérius O. Bonates, and Carlile Lavor

**BigData Industry and Application Session 9 - Hadoop**
Big R: Large-Scale Analytics on Hadoop Using R
Oscar D. Lara Yejas, Weiqiang Zhuang, and Adarsh Pannu
Diagnosing Virtualized Hadoop Performance from Benchmark Results: An Exploratory Study
Jun Fan, Xinhui Li, Chi Harold Liu, Jeffrey Buell, Gavin Lu, and Luke Lu
Performance Implications of SSDs in Virtualized Hadoop Clusters
Sungyong Ahn, Sangkyu Park, Jae-Ki Hong, and Wooseok Chang

**BigData Industry and Application Session 10 - BigData Privacy**
We Know Where You Are Tweeting From: Assigning a Type of Place to Tweets Using Natural Language Processing and Random Forests
Abdulkareem Alsudais, Gondy Leroy, and Anthony Corso
Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues
Jeff Sedayao, Rahul Bhardwaj, and Nakul Gorade

**BigData Industry and Application Session 11 - Cloud-Based BigData Storage**
CloudJet4BigData: Streamlining Big Data via an Accelerated Socket Interface
Frank Zhigang Wang, Theo Dimitrakos, Na Helian, Sining Wu, Ling Li, and Rodric Yates
Resource Provisioning with QoS in Cloud Storage
Wei-Chih Huang, Chuan-Ming Liu, and Chuan-Chi Lai
Data Sharing for Cloud Computing Platforms
Marwan Sabbouh, Kenneth McCracken, and Geoff Cooney
## BigData Industry and Application Session 12 - BigData Analytics Architecture

Towards a Set Theoretical Approach to Big Data Analytics ..........................................................629  
  *Raghava Rao Mukkamala, Abid Hussain, and Ravi Vatrapu*

Efficient String Sorting on Multi - and Many-Core Architectures ...................................................637  
  *Aleksandr Drozd, Miquel Pericàs, and Satoshi Matsuoka*

A Tale of Two Data-Intensive Paradigms: Applications, Abstractions, and Architectures ........................................645  
  *Shantenu Jha, Judy Qiu, Andre Luckow, Pradeep Mantha, and Geoffrey C. Fox*

## BigData Industry and Application Session 13 - NoSQL

ODBAPI: A Unified REST API for Relational and NoSQL Data Stores ........................................653  
  *Rami Sellami, Sami Bhiri, and Bruno Defude*

Terms Mining in Document-Based NoSQL: Response to Unstructured Data ..................................661  
  *Richard K. Lomotey and Ralph Deters*

Spatio-temporal Pseudo Relevance Feedback for Large-Scale and Heterogeneous Scientific Repositories ..................................................669  
  *Shin'ichi Takeuchi, Yuhei Akahoshi, Bun Theang Ong, Komei Sugiura, and Koji Zettsu*

An Intelligent System for Forecasting the Trend of Consumed Electricity ......................................677  
  *Hang Yang, Huajun Chen, Cai Yuan, and Fang Lianhang*

## BigData Industry and Application Session 14 - NoSQL

Tactical Clinical Text Mining for Improved Patient Characterization ...........................................683  
  *Heather Champion, Nick Pizzi, and Raja Krishnamoorthy*

An Effective Integrated Method for Learning Big Imbalanced Data ..............................................691  
  *Mojgan Ghanavati, Raymond K. Wong, Fang Chen, Yang Wang, and Chang-Shing Perng*

## BigData Industry and Application Session 15 - Big Social Data

Multifaceted Visualisation of Annotated Social Media Data .......................................................699  
  *Sanat Kumar Bista, Surya Nepal, and Cécile Paris*
BigData Industry and Application Session 16 - Graph Analytics
Enumerating Maximal Bicliques from a Large Graph Using MapReduce ....................................................707
Arko Provo Mukherjee and Srikanta Tirthapura

LightGraph: Lighten Communication in Distributed Graph-Parallel Processing ..................................................717
Yue Zhao, Kenji Yoshigoe, Mengjun Xie, Suijian Zhou, Remzi Seker, and Jiang Bian

RABID: A Distributed Parallel R for Large Datasets ..........................................................................................................................725
Hao Lin, Shuo Yang, and Samuel P. Midkiff

BigData Industry and Application Session 17 - Cloud-Based BigData Service
Trace-Based Analysis and Prediction of Cloud Computing User Behavior Using the Fractal Modeling Technique ..........................................................733
Shuang Chen, Mahboobeh Ghorbani, Yanzhi Wang, Paul Bogdan, and Massoud Pedram

Inherent Replica Inconsistency in Cassandra .................................................................................................................................740
Xiangdong Huang, Jianmin Wang, Jian Bai, Guiguang Ding, and Mingsheng Long

Towards Emulation of Large Scale Complex Network Workloads on Graph Databases with XGDBench .................................................................748
Miyuru Dayarathna and Toyotaro Suzumura

BigData Industry and Application Session 18 - Data Cleaning Model and Performance

BigData Work-in-Progress Session 1 - BigData Processing
Policy Design Based on Risk at Big Data Era: Case Study of Privacy Invasion in South Korea .................................................................756
Hyejung Moon, Hyun Suk Cho, Seo Hwa Jeong, and Jangho Park

Knowledge Discovery from Big Data for Intrusion Detection Using LDA ............................................................................................................................760
Jingwei Huang, Zbigniew Kalbarczyk, and David M. Nicol

Big Data Security and Privacy Issues in Healthcare ............................................................................................................................762
Harsh Kupwade Patil and Ravi Seshadri

GrandLand Traffic Data Processing Platform ...............................................................................................................................766
Xingcan Cui, Zhen Dong, Liwei Lin, Renyong Song, and Xiaohui Yu

A Memory-Based Continuous Query Index for Stream Processing ........................................................................................................768
Cuiwen Xiong, Peng Zhang, Yan Li, Shipeng Zhang, Qingyun Liu, and Jianlong Tan
BigData Work-in-Progress Session 2 - BigData Analytics

Ranking Keyword Search Results with Query Logs ................................................................. 770
  Jing Zhou, Xiaohui Yu, Yang Liu, and Ziqiang Yu

Constructing an Issue Network from the Perspective of Common R&D Keywords ................................................................................................................................. 772
  Namgyu Kim, William Wong Xiu Shun, Jieun Kim, Kee-Young Kwahk,
  Seungryul Jeong, and Hyunchul Ahn

Agriculture Data for All - Integrated Tools for Agriculture Data Integration,
Analytics, and Sharing ........................................................................................................... 774
  Jarek Nabrzyski, Cheng Liu, Charles Vardeman, Sandra Gesing,
  and Milan Budhatoki

Asynchronous Service Analysis of Cloud DVR DataCenter ............................................. 776
  Zhen Zhao

Work in Progress - In-Memory Analysis for Healthcare Big Data ........................................ 778
  Muaz Mian, Ankur Teredesai, David Hazel, Sreenivasulu Pokuri,
  and Krishna Uppala

Large-Scale Heterogeneous Program Retrieval through Frequent Pattern Discovery and Feature Correlation Analysis ................................................................. 780
  Bo Liu, Liang Wu, Qiuxiang Dong, and Yuanchun Zhou

BigData Work-in-Progress Session 3 - BigData Management Framework

Towards Efficient KNN Joins on Data Streams ................................................................. 782
  Chong Yang, Xiaohui Yu, and Yang Liu

CouchFS: A High-Performance File System for Large Data Sets ........................................ 784
  Fangzhou Yao and Roy H. Campbell

A Computational Framework for Integrating and Retrieving Biodiversity Data on a Large Scale ................................................................................................................... 786
  Daniel L. Da Silva, Pedro L.P. Corrêa, Silvio L. Stanzani, Paulo André Filipak,
  and Andreiwid Sheffer Corrêa

Versioning Complex Data .................................................................................................... 788
  Matt Macduff, Benno Lee, and Sherman Beus

On the Way to Big Data Applications in Industrial Computed Tomography ....................... 792
  Alexander Ditter, Dietmar Fey, Tobias Schön, and Steven Oeckl
BigData Work-in-Progress Session 4 - BigData Decision Support System

Towards a Novel Model for Distributed Big Data Service Composition Using Functional Graph Matching ................................................................. 794
   Carlos R. Rivero and Hasan M. Jamil

Enrichment Patterns for Big Data ........................................................................................................................................... 796
   Kerrie Holley, Gandhi Sivakumar, and Kalapriya Kannan

YCSB and TPC-H: Big Data and Decision Support Benchmarks ................................................................. 800
   Melyssa Barata, Jorge Bernardino, and Pedro Furtado

Understanding Student Behaviors in Online Classroom: Data Scientific Approach ........................................................................................................................................ 802
   Jongbok Byun, Diane Pennington, Jorge Cardenas, Srabasti Dutta,
   and Jeral Kirwan

Scientific Data Infrastructure for Sustainability Science Mobile Applications ................................................. 804
   Katherine G. Herbert, Emily Hill, Jerry Alan Falls, Joseph O. Ajala,
   Richard T. Boniface, and Paul W. Cushman

Really Opened Government Data: A Collaborative Transparency at Sight ......................................................... 806
   Andreiwid Sheffer Corrêa, Pedro L.P. Corrêa, Daniel L. Silva,
   and Flávio Soares Corrêa Da Silva

Classifying and Aggregating Context Attributes for Business Service Requests
- No 'One-Size-Fits-All' ........................................................................................................................................ ................. 808
   Rebecca Copeland and Noel Crespi

Big Data Infrastructure for analyzing data generated by Wireless Sensor Networks ................................................................. 816
   Lidice Garcia Rios and Jos’e Alberto Incera Diguez

Automated Trading with Machine Learning on Big Data ......................................................................................... 824
   Dymitr Ruta

Author Index ......................................................................................................................................................... 831