

2018 IEEE International Congress on Big Data (BigData Congress 2018)

**San Francisco, California, USA
2-7 July 2018**



**IEEE Catalog Number: CFP18SEV-POD
ISBN: 978-1-5386-7233-4**

**Copyright © 2018 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 is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP18SEV-POD
ISBN (Print-On-Demand):	978-1-5386-7233-4
ISBN (Online):	978-1-5386-7232-7
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

2018 IEEE International Congress on Big Data **BigDataCongress 2018**

Table of Contents

Message from the IEEE BigData Congress 2018 Chairs .xi.....
IEEE BigData Congress 2018 Organizing Committee .xii.....
IEEE BigData Congress 2018 Reviewers .xix.....

Regular Papers

Session 1: Big Data Models and Algorithms

Compile-Time Code Generation for Embedded Data-Intensive Query Languages .1.....
*Leonidas Fegaras (University of Texas at Arlington) and Md
Hasanuzzaman Noor (University of Texas at Arlington)*

Lambda-Blocks: Data Processing with Topologies of Blocks .9.....
*Matthieu Caneill (Université Grenoble Alpes) and Noël De Palma
(Université Grenoble Alpes)*

Incorporating Word Embedding into Cross-Lingual Topic Modeling .17.....
*Chia-Hsuan Chang (National Sun Yat-sen University), San-Yih Hwang
(National Sun Yat-sen University), and Tou-Hsiang Xui (National Sun
Yat-sen University)*

Session 2: Big Data Mining and Visualization

On the Usage of the Probability Integral Transform to Reduce the Complexity of Multi-Way
Fuzzy Decision Trees in Big Data Classification Problems .25.....
*Mikel Elcano (Public University of Navarre), Mikel Uriz (Public
University of Navarre), Humberto Bustince (Public University of
Navarre), and Mikel Galar (Public University of Navarre)*

Useful ToPIC: Self-Tuning Strategies to Enhance Latent Dirichlet Allocation .33.....
*Stefano Proto (Politecnico di Torino), Evelina Di Corso (Politecnico
di Torino), Francesco Ventura (Politecnico di Torino), and Tania
Cerquitelli (Politecnico di Torino)*

A Survey of Current End-User Data Analytics Tool Support .41.....
*Hourieh Khalajzadeh (Deakin University), Mohamed Abdelrazek (Deakin
University), John Grundy (Monash University), John Hosking (University
of Auckland), and Qiang He (Swinburne University of Technology)*

Session 3: Big Data – Smart Cities and IoT

- Treepedia 2.0: Applying Deep Learning for Large-Scale Quantification of Urban Tree Cover .49.....
Bill Yang Cai (Massachusetts Institute of Technology), Xiaojiang Li (Massachusetts Institute of Technology), Ian Seiferling (Massachusetts Institute of Technology), and Carlo Ratti (Massachusetts Institute of Technology)
- Short-Term Traffic Prediction Using Long Short-Term Memory Neural Networks .57.....
Zainab Abbas (KTH Royal Institute of Technology), Ahmad Al-Shishtawy (RISE SICS), Sarunas Girdzijauskas (KTH Royal Institute of Technology), and Vladimir Vlassov (KTH Royal Institute of Technology)
- A Data-Driven Approach to Predict an Individual Customer's Call Arrival in Multichannel Customer Support Centers .66.....
Somayeh Moazeni (Stevens Institute of Technology) and Rodrigo Andrade (Stevens Institute of Technology)
- Analysing Customer Engagement of Turkish Airlines Using Big Social Data .74.....
Fie Sternberg (Copenhagen Business School), Kasper Hedegaard Pedersen (Copenhagen Business School), Niklas Klve Ryelund (Copenhagen Business School), Raghava Rao Mukkamala (Copenhagen Business School), and Ravi Vatraru (Copenhagen Business School)

Session 4: Big Data – Health and Applications

- Diagnosis Recommendation Using Machine Learning Scientific Workflows .82.....
Ishtiaq Ahmed (Wayne State University), Shiyong Lu (Wayne State University), Changxin Bai (Wayne State University), and Fahima Amin Bhuyan (Wayne State University)
- Nowcasting Events from Twitter Social Media with Semi-Supervised Learning .91.....
Jin Soung Yoo (Purdue University Fort Wayne) and David Kimmey (Purdue University Fort Wayne)
- Big Web Colors: Analyzing the World Top Sites .96.....
Massimo Marchiori (University of Padua) and Giulio Rigoni (University of Padua)

Session 5: Big Data Management

- Towards a Better Replica Management for Hadoop Distributed File System .104.....
Hilmi Egemen Ciritoglu (University College Dublin), Takfarinas Saber (University College Dublin), Teodora Sandra Buda (IBM Ireland), John Murphy (University College Dublin), and Christina Thorpe (University College Dublin)
- Budget-Transfer: A Low Cost Inter-Service Data Storage and Transfer Scheme .112.....
Galen Deal (University of Washington Bothell), Yang Peng (University of Washington Bothell), and Hua Qin (Hunan City University)
- GDedup: Distributed File System Level Deduplication for Genomic Big Data .120.....
Paul Bartus (University of Puerto Rico, Mayaguez Campus) and Emmanuel Arzuaga (University of Puerto Rico, Mayaguez Campus)

Session 6: Big Data Analytics

- A Fourier-Based Data Minimization Algorithm for Fast and Secure Transfer of Big Genomic Datasets .128.....
Mohammed Aledhari (Western Michigan University), Marianne Di Pierro (Western Michigan University), and Fahad Saeed (Western Michigan University)
- Performance Modeling and Task Scheduling in Distributed Graph Processing .135.....
Daniel Presser (Universidade Federal de Santa Catarina), Frank Siqueira (Universidade Federal de Santa Catarina), and Fabio Reina (Universidade Federal de Santa Catarina)
- Exploration of Bi-Level PageRank Algorithm for Power Flow Analysis Using Graph Database .143....
Chen Yuan (Global Energy Interconnection Research Institute North America), Yi Lu (State Grid Sichuan Electric Power Company), Kewen Liu (Global Energy Interconnection Research Institute), Guangyi Liu (Global Energy Interconnection Research Institute North America), Renchang Dai (Global Energy Interconnection Research Institute North America), and Zhiwei Wang (Global Energy Interconnection Research Institute North America)

Session 7: Architecture Solution and Quality of Big Data Service

- Dynamic Model Evaluation to Accelerate Distributed Machine Learning .150.....
Simon Caton (National College of Ireland), Srikumar Venugopal (IBM Research), Shashi Bhushan TN (National College of Ireland), Vidya Sankar Velamuri (National College of Ireland), and Kostas Katrinis (IBM Research)
- Sensor Data Based System-Level Anomaly Prediction for Smart Manufacturing .158.....
Jianwu Wang (University of Maryland, Baltimore County), Chen Liu (North China University of Technology), Meiling Zhu (North China University of Technology), Pei Guo (University of Maryland, Baltimore County), and Yapeng Hu (North China University of Technology)
- Big Data Quality: A Survey .166.....
Ikbal Taleb (Concordia University), Mohamed Adel Serhani (United Arab Emirates University), and Rachida Dssouli (Concordia University)

Workshop Papers

- A Fast and Incremental Development Life Cycle for Data Analytics as a Service .174.....
Claudio A. Ardagna (Università degli Studi di Milano), Valerio Bellandi (Università degli Studi di Milano), Paolo Ceravolo (Università degli Studi di Milano), Ernesto Damiani (Università degli Studi di Milano), Beniamino Di Martino (Università degli Studi della Campania), Salvatore D'Angelo (Università degli Studi della Campania), and Antonio Esposito (Università degli Studi della Campania)

XRT: Programming-Language Independent MapReduce on Shared-Memory Systems .182.....	<i>Erik Selin (University of Ottawa) and Herna Viktor (University of Ottawa)</i>
An Architecture for Cost Optimization in the Processing of Big Geospatial Data in Public Cloud Providers .190.....	<i>Joao Bachiega Jr. (University of Brasilia), Marco Antonio Sousa Reis (University of Brasilia), Maristela Holanda (University of Brasilia), and Aleteia P. F. Araujo (University of Brasilia)</i>
Estimation of Types of States in Partial Observable Network Systems .198.....	<i>Sayantan Guha (Arizona State University)</i>

Work-in-Progress Papers

Session 1

Autoencoder Evaluation and Hyper-Parameter Tuning in an Unsupervised Setting .205.....	<i>Ellie Ordway-West (AT&T), Pallabi Parveen (AT&T), and Austin Henslee (AT&T)</i>
Learning a Joint Low-Rank and Gaussian Model in Matrix Completion with Spectral Regularization and Expectation Maximization Algorithm .210.....	<i>Gang Wu (Iowa State University) and Ratnesh Kumar (Iowa State University)</i>
DynMDL: A Parallel Trajectory Segmentation Algorithm .215.....	<i>Eleazar Leal (University of Minnesota Duluth) and Le Gruenwald (University of Oklahoma)</i>
Insights on Apache Spark Usage by Mining Stack Overflow Questions .219.....	<i>Leonardo Jiménez Rodríguez (Samsung Research America), Xiaoran Wang (Samsung Research America), and Jilong Kuang (Samsung Research America)</i>
Biparti Majority Learning with Tensors .224.....	<i>Chia-Lun Lee (National Chengchi University), Shun-Wen Hsiao (National Chengchi University), and Fang Yu (National Chengchi University)</i>

Session 2

An OWL Ontology for Supporting Semantic Services in Big Data Platforms .228.....	<i>Domenico Redavid (Consorzio Interuniversitario Nazionale per l'Informatica), Roberto Corizzo (Consorzio Interuniversitario Nazionale per l'Informatica), and Donato Malerba (University of Bari Aldo Moro)</i>
Graph-Based Data Relevance Estimation for Large Storage Systems .232.....	<i>Vinodh Venkatesan (IBM Research), Taras Lehinevych (National University of Kyiv-Mohyla Academy, Kiev), Giovanni Cherubini (IBM Research), Andrii Glybovets (National University of Kyiv-Mohyla Academy, Kiev), and Mark Lantz (IBM Research)</i>

BigDataStack: A Holistic Data-Driven Stack for Big Data Applications and Operations .237.....
Dimosthenis Kyriazis (University of Piraeus), Christos Doulkeridis (University of Piraeus), Panagiotis Gouvas (Ubitech), Ricardo Jimenez-Peris (LeanXcale), Ana Juan Ferrer (Atos Research and Innovation), Leonidas Kallipolitis (Athens Technology Center), Pavlos Kranas (LeanXcale), George Kousiouris (University of Piraeus), Craig Macdonald (University of Glasgow), Richard McCreddie (University of Glasgow), Yosef Moatti (IBM Research), Apostolos Papageorgiou (NEC Laboratories Europe), Marta Patino-Martinez (Universidad Politecnica de Madrid), Stathis Plitsos (Danaos Shipping), Dimitris Pouloupoulos (University of Piraeus), Antonio Paradell (Atos Research and Innovation), Amaryllis Raouzaïou (Athens Technology Center), Paula Ta-Shma (IBM Research), and Valerio Vianello (Universidad Politecnica de Madrid)

Stream Analytics and Adaptive Windows for Operational Mode Identification of Time-Varying Industrial Systems .242.....
Athar Khodabakhsh (Ozyegin University), Ismail Ari (Ozyegin University), Mustafa Bakir (TUPRAS), and Serhat Murat Alagoz (TUPRAS)

Latency Measurement of Fine-Grained Operations in Benchmarking Distributed Stream Processing Frameworks .247.....
Giselle van Dongen (Ghent University), Bram Steurtewagen (Ghent University), and Dirk Van den Poel (Ghent University)

Session 3

Large Scale Predictive Analytics for Hard Disk Remaining Useful Life Estimation .251.....
Preethi Anantharaman (IBM Almaden Research Center), Mu Qiao (IBM Almaden Research Center), and Divyesh Jadav (IBM Almaden Research Center)

Adaptive Trip Recommendation System: Balancing Travelers among POIs with MapReduce .255.....
Sara Migliorini (University of Verona), Damiano Carra (University of Verona), and Alberto Belussi (University of Verona)

A Personalized Travel Recommendation System Using Social Media Analysis .260.....
Joseph Coelho (Marquette University), Paromita Nitu (Marquette University), and Praveen Madiraju (Marquette University)

Time Series Sanitization with Metric-Based Privacy .264.....
Liyue Fan (University at Albany, SUNY) and Luca Bonomi (University of California, San Diego)

Towards Optimal Snapshot Materialization to Support Large Query Workload for Append-Only Temporal Databases .268.....
Amin Beiraimi (University of Ontario Institute of Technology), Ken Pu (University of Ontario Institute of Technology), and Ying Zhu (University of Ontario Institute of Technology)

Categorical Models for BigData .272.....
Laurent Thiry (IRIMAS), Heng Zhao (IRIMAS-Universite de Haute Alsace), and Michel Hassenforder (IRIMAS-Universite de Haute Alsace)

Author Index 277.....