Editorial Preface: Special Issue on Big Data Analytics, Infrastructure, and Applications

S. Sobolevsky, S. McIntosh, and P.C.K. Hung

THE pervasive nature of digital technologies as witnessed in industry, services and everyday life has given rise to an emergent, data-focused economy stemming from many aspects of human individual and commercial activity. The richness and vastness of these data are creating unprecedented research opportunities in a number of fields including urban studies, geography, economics, finance, and social science, as well as physics, biology and genetics, public health and many others. In addition to Big Data-inspired research, businesses have seized on big data technologies to support and propel growing business intelligence needs. As businesses build out Big Data hardware and software infrastructure, it becomes increasingly important to anticipate technical and practical challenges and to identify best practices learned through experience. Big Data analytics employ software tools from advanced analytics disciplines such as data mining, predictive analytics, and machine learning. At the same time, the processing and analysis of big data presents methodological and technological challenges. The goal of this special issue is to present both novel solutions to challenging technical issues as well as compelling Big Data use cases.

This special issue contains 13 papers that provide deep research results to report the advance of *Big Data Analytics, Infrastructure, and Applications*. The first contribution is "Dynamic Job Ordering and Slot Configurations for MapReduce Workloads" by Tang et al. This paper presents two algorithms to minimize the make span and completion time for offline MapReduce workloads by job ordering optimization under a given map/reduce slot configuration. The paper also shows experimental results through simulations on Amazon EC2 to show significant reductions in running time in practice. Next the second paper "Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-mining-as-a-service Paradigm" by Dong et al. presents probabilistic and deterministic verification methods to measure the result correctness and completion of frequent item sets from a cloud data server. This paper also demonstrates the effectiveness and efficiency of the proposed methods by using a set of empirical results based on real datasets.

- *S. Sobolevsky is with MIT, Cambridge, MA. E-mail: stanly@mit.edu.*
- *S. McIntosh is with Cloudera Inc. and New York University. E-mail: mcintosh@cs.nyu.edu.*
- *P.C.K. Hung is with the National Taipei University of Technology, Taipei, Taiwan and University of Ontario Institute of Technology, Oshawa, ON, Canada. E-mail: patrick.hung@uoit.ca.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TSC.2015.2509738

Then Fong et al. present a lightweight feature selection for mining real time streaming data based on the high-dimensionality and streaming format of data feeds by using accelerated particle swarm optimization (APSO). This paper "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data" also addresses an experiment result by a collection of Big Data under the proposed feature selection algorithm for performance evaluation.

Referring to the forth paper "Processing Cassandra Datasets with Hadoop-Streaming Based Approaches," Bedri et al. present an approach for integrating NoSQL data stores with MapReduce under non-Java application scenarios. This paper compares the approach to Hadoop Streaming on file-system based data stores and the result shows the effectiveness of the proposed approach. Further the fifth paper "Supporting multi data stores applications in cloud environments" by Bhiri et al. discusses a unifying data model to interact with virtual relational and NoSQL data stores by using OPEN-PaaS-DataBase API (ODBAPI) queries and REST API allowing the programming of applications independently of the target databases. The paper presents a declarative approach to lighten the burden of the tedious tasks of discovering data stores and deploying applications in cloud environments with a prototype. Next Zhang et al. present an integrated optimal localized skyline query processing method for building up cloud mashup applications. This paper "Skyline Discovery and Composition of Inter-Cloud Mashup Services" shows effective experimental results on the Quality of Web Service (QWS) benchmark data along six QoS dimensions. Then the paper "BeTL: MapReduce Checkpoint Tactics Beneath the Task Level" by Chen et al. presents a tactic named BeTL with slight changes to the execution flow of MapReduce to achieve a finer-grained fault tolerance. This paper also shows experimental results involved less IO operations and performed better than Hadoop even under no failures.

The next contribution is "Dynamic Virtual Chunks: On Supporting Efficient Accesses to Compressed Scientific Data" by Zhao et al. addresses logical blocks in virtual chunks with efficient random accesses to the compressed scientific data without sacrificing the compression ratio and the physical continuity of the file content. This paper discusses the algorithms, analysis, and evaluations of dynamic virtual chunks to deal with cases where the references are updated dynamically. Referring to the ninth paper "Distributed Maximal Clique Computation and Management," Cheng et al. present a distributed query operation and efficient algorithm built on a shared-nothing architecture for computing the set of maximal

cliques in common real-world graphs. This paper verifies the efficiency of the algorithms for computing, querying, and updating the set of maximal cliques with a range of real-world graphs from different application domains. Next Ardagna et al. present a score-based benchmark for NoSQL databases, which is independent from the specific configurations of the database and deployment environment. This paper "A Configuration-Independent Score-Based Benchmark for Distributed Databases" also evaluates databases according to different properties such as performance and consistency, which can be integrated with existing benchmarks to reduce the burden of their execution.

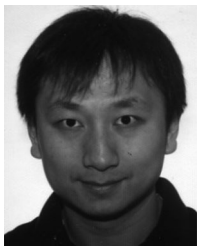
The next paper "Towards achieving Data Security with the Cloud Computing Adoption Framework" by Chang and Ramachandran presents a multi-layered model called Cloud Computing Adoption Framework (CCAF) for Big Data security protection. This paper also uses Business Process Modeling Notation (BPMN) to simulate the proposed model with experimental results. Then the paper "Common Pitfalls of Benchmarking Big Data Systems" by Chen and Shapira discuss five common pitfalls, namely "Comparing Apples to Orange," "Not Testing at Scale," "Believing in Miracles," "Using Unrealistic Benchmarks," and "Communicating Results Poorly," drawn from engineering and customer experiences at Cloudera. This paper offers a behind-the-scenes at internal engineering and review processes that produces rigorous benchmark results. Lastly Yang et al. discuss a deep computation model for feature learning on Big Data by using a high-order back-propagation algorithm (HBP) by extending the conventional back-propagation algorithm from the vector space to the high-order tensor space. This paper "Deep Computation Model for Unsupervised Feature Learning on Big Data" also shows experimental results on representative datasets by comparison with stacking auto-encoders and multimodal deep learning models.

In summary, the goal of this special issue is to crystallize the emerging Big Data technologies and trends into positive efforts to focus on the most promising solutions in industry. The papers provide clear proof that Big Data technologies are playing a more and more important and critical role in supporting various applications in industry. It is also believed that the papers will further research new best practices and directions in this emerging research discipline.

Stanislav Sobolevsky
Suzanne McIntosh
Patrick C.K. Hung
Guest Editors

Stanislav Sobolevsky received the PhD degree in 1999 and doctor of science (habilitation degree) in 2009 in mathematics and applies his fundamental quantitative background to studying human behavior in urban context through its digital traces—spatio-temporal big data created by various aspects of human activity. He has been an Associate Professor of Practice at the Center for Urban Science and progress at New York University since 2015. His research interests cover network science, big data analytics, modeling of complex systems, and the theory of differential equations. He authored one monograph, two textbooks and over 60 peer-reviewed papers in mathematics, network science, and mathematical modeling. His former work experience includes research, professorship, and administrative positions at MIT, Belarusian State University and Academy of Science of Belarus. Served as a guest editor of the special issue on Big Data Analytics, Infrastructure and Applications of *IEEE Transactions on Services Computing*, one of the organizers of IEEE Big Data Congress' 2015 and the symposium "CompleXity: Big Data and Technology for Complex Urban Systems" at The Hawaii International Conference on System Sciences (HICSS) in January, 2016, program committee member at a number of a number of IEEE and ACM computer science conferences.

Suzanne McIntosh received the Engineer Degree in Computer Engineering from Stevens Institute of Technology. She is an adjunct professor at New York University Courant Institute of Mathematical Sciences, Center for Data Science, and Tandon School of Engineering. She is also a technology consultant with Cloudera and previously was with IBM T.J. Watson Research Center in Yorktown Heights, N.Y., where she led cross-disciplinary research teams in virtualization, analytics technologies, and security research. She is an inventor with patents in security and virtualization technologies for which she was awarded two IBM Research Invention Plateau awards and two IBM Research Technical Accomplishment Awards for Open Source Contributions to Virtualization Security, and for the Caernarvon Secure Operating System. Prior to IBM Research, she developed secure, real-time-embedded systems for battlefield communications and GPS satellites. She is a guest editor for *Transactions on Services Computing Special Issue on Big Data Analytics, Infrastructure, and Applications*, program and panels chair for several Big Data and security conferences including International Congress on Big Data and ACSAC, membership chair for ACM SIGHPC, and founder and Member-At-Large of ACM SIGHPC-BigData. She is a senior member of the ACM and IEEE, and she serves as a mentor through the Society of Women Engineers mentoring program.



Patrick C.K. Hung is an Associate Professor at the Faculty of Business and Information Technology in University of Ontario Institute of Technology, Canada. He has been working with Boeing Research and Technology on aviation services-related research with two U.S. patents on mobile network dynamic workflow system. He is an Honorary International Chair Professor at National Taipei University of Technology in Taiwan, an Adjunct Professor at Nanjing University of Information Science & Technology in China, and a Visiting Professor at University of Sao Paulo in Brazil. In addition, he was an Adjunct Professor at Wuhan University, a Visiting Professor at the Shizuoka University and University of Aizu in Japan, a Guest Professor in University of Innsbruck in Austria, University of Trento and University of Milan in Italy. Before that, he was a Research Scientist with Commonwealth Scientific and Industrial Research Organization in Australia as well as he worked as a software engineer in industry in North America. He is a founding committee member of the IEEE International Conference of Web Services (ICWS), IEEE International Conference on Services Computing (SCC), and IEEE BigData Congress (BigData Congress). He is also an Associate Editor of the *IEEE Transactions on Services Computing*, and a Coordinating Editor of the Information Systems Frontiers.