Guest Editorial: Big Data Infrastructure I

Jinjun Chen and Honggang Wang

DATA is becoming an increasingly decisive resource in modern societies, economies, and governmental organizations. Big Data is an emerging paradigm encompassing various kinds of complex and large scale information beyond the processing capability of conventional software and databases. Various technologies are being discussed to support the handling of big data such as massively parallel processing databases, scalable storage systems, cloud computing platforms, Hadoop and Spark. Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, large-scale scalable Big Data Infrastructure with corresponding programming language support and software models for efficient processing in distributed environments such as cloud is on demand.

In this special issue, we invite articles on innovative research to address challenges of Big Data Infrastructure with emerging computing platforms such as heterogeneous clouds, hybrid architectures, Hadoop or Spark with emphasis on addressing real-time requirements imposed by emerging Big Data applications such as sensing data, e-commerce data, business transactions and web logs, and etc.

It is a great honor of the editorial team to have more than 80 research groups share their latest research works and solutions to these challenges. The SI attracted a large number of submissions from North America, Europe, Australia, and Asia. Our Guest Editor (GE) team accepted 21 papers that cover multiple aspects of Big Data Infrastructure. We hope that this SI will serve as a good reference for engineers, scientists, researchers, and academics in the field of Big Data. We grouped the accepted papers into two editorials: Big Data Infrastructure I and II. This is editorial for Big Data Infrastructure I.

In the SI, the first paper titled “HDM: A Composable Framework for Big Data Processing” presents the Hierarchically Distributed Data Matrix (HDM) which is a functional, strongly-typed data representation for writing composable big data applications; In the second paper titled “Distributed Feature Selection for Efficient Economic Big Data Analysis”, the authors propose a framework that combines the methods of economic feature selection and econometric model construction to reveal the hidden patterns for economic development; In the third paper titled “Thermal-Aware and DVFS-Enabled Big Data Task Scheduling for Data Centers”, the authors aim to achieve maximum energy efficiency by combining thermal-aware and dynamic voltage and frequency scaling (DVFS) techniques. The fourth paper is titled as “GPU-Accelerated High-Throughput Online Stream Data Processing” which presents the design, implementation and evaluation of G-Storm, a GPU-enabled parallel system based on Storm; In the fifth paper titled “MetaFlow: A Scalable Metadata Lookup Service for Distributed File Systems in Data Centers”, the author present MetaFlow, a scalable metadata lookup service utilizing software-defined networking (SDN) techniques to distribute lookup workload over network components; The sixth paper titled “Complex Queries Optimization and Evaluation over Relational and NoSQL Data Stores in Cloud Environments” presents a mediation based component to optimize and execute complex queries over multiple data stores in Cloud environments. In the Seventh paper titled “Speed Up Big Data Analytics by Unveiling the Storage Distribution of Sub-Datasets”, an efficient algorithm to obtain the meta-data of sub-dataset distributions is proposed. In addition, the authors design an elastic storage structure called ElasticMap based on the HashMap and BloomFilter techniques to store the meta-data. The major contribution in the eighth paper titled “From Latency, Through Outbreak, to Decline: Detecting Different States of Emergency Events Using Web Resources” includes the development of web resources based states detecting algorithm of an event and discovery on the relationship between web and emergency events; In the ninth paper titled “JouleMR: Towards Cost-Effective and Green-Aware Data Processing Frameworks”, the authors propose JouleMR, a cost-effective and green-aware data processing framework; The tenth paper titled “CaL: Extending Data Locality to Consider Concurrency for Performance Optimization” introduces the concept and definition of Concurrency-aware data access Locality (CaL), which, as its name states, extends the concept of locality by considering concurrency. In the 11th paper titled “An Enhanced Visualization Method to Aid Behavioral Trajectory Pattern Recognition Infrastructure for Big Longitudinal Data”, the authors propose an enhanced projection pursuit (EPP) method to better project and visualize the structures (e.g., clusters) of big high-dimensional (HD) longitudinal data on a lower-dimensional plane.

We would like to thank all the authors for their contributions and the reviewers for their great effort in the SI review process. We would especially like to thank the Editor-in-Chief, Prof. Qiang Yang for allowing our GE team to...
have the opportunity to publish this SI and his great support. We would also like to thank Ms. Kathy Santa Maria for her dedication and editorial support for the SI. We believe that the papers from the SI will contribute to the further development of big data infrastructure and applications and inspire future research in the field.

Jinjun Chen
Honggang Wang
Guest Editor

Jinjun Chen received the PhD degree in information technology from the Swinburne University of Technology, Australia. He is a professor from the Swinburne University of Technology, Australia. He is the deputy director of Swinburne Data Science Research Institute, and director of Swinburne Big Data Lab. His research interests include scalability, big data, data science, software systems, cloud computing, data privacy and security, and related various research topics. His research results have been published in more than 130 papers in international journals and conferences, including various IEEE/ACM Transactions. He received UTS Vice-Chancellor’s Awards for Research Excellence Highly Commended (2014), UTS Vice-Chancellor’s Awards for Research Excellence Finalist (2013), Swinburne Vice-Chancellors Research Award (ECR) (2008), IEEE Computer Society Outstanding Leadership Award (2008-2009) and (2010-2011), IEEE Computer Society Service Award (2007), Swinburne Faculty of ICT Research Thesis Excellence Award (2007). He is an associate editor of the ACM Computing Surveys, the IEEE Transactions on Big Data, the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Cloud Computing, as well as other journals such as the Journal of Computer and System Sciences, the Journal of Network and Computer Applications. He is the chair of the IEEE Computer Society’s Technical Committee on Scalable Computing (TCSC).

Honggang Wang received the PhD degree in computer engineering from the University of Nebraska-Lincoln, in 2009. He is the “Scholar of The Year” (2016, the highest research recognition at UMass Dartmouth). Before he joined UMass Dartmouth in 2009, he has worked for Bell Labs Lucent Technologies China from 2001 to 2004 as a member of Technical Staff. He is a tenured associate professor at UMass Dartmouth and is an affiliated faculty member of Advanced Telecommunications Engineering Laboratory, University of Nebraska-Lincoln. He is also the faculty member of Biomedical Engineering and Biotechnology PhD program (BMEBT), UMass Dartmouth. His research interests include internet of things, wireless health, body area networks (BAN), cyber and multimedia security, mobile multimedia and cloud, wireless networks and cyber-physical system, and BIG DATA in mHealth. He has published more than 150 papers in his research areas. He was an invited participant by National Academic Engineering (NAE) for 2017 German-American Frontiers of Engineering Symposium, as one of about 50 outstanding engineers (ages 30-45) from US companies, universities, and government labs. He has served as the general chair/co-chair and TPC chair/co-chairs for many IEEE conferences. He also serves as the steering committee co-chair of IEEE CHASE and TPC co-chair of IEEE CHASE 2016, which is a leading international conference in the field of connected health. He has also been serving as the associate editor in chief (EiC) of the IEEE Internet of Things Journal (SCI impact factor: 7.596), associate editors of the IEEE Transactions on Big Data, the IEEE Transactions on Multimedia, and the IEEE Access, an editor of the IEEE Transactions on Vehicular Technology, associate technical editor of the IEEE Communication Magazine. He received IEEE Multimedia Communications Technical Committee (MMTC) Outstanding Leadership Award (2015) and IEEE HEALTHCOM 2015 Outstanding Service Award. His research is supported by NSF DoT, UMass President office, and UMass Healey Grant (total over 2M dollars).