Digital Object Identifier 10.1109/TETC.2016.2524219

Advances in Methods and Techniques for Processing Streaming Big Data in Datacentre Clouds

INTERNET of Things (IoT), a part of Future Internet, comprises many billions of Internet connected Objects (ICOs) or 'things' where things can sense, communicate, compute and potentially actuate as well as have intelligence, multi-modal interfaces, physical/ virtual identities and attributes. The IoT vision has recently given rise to emerging IoT big data applications [2], [3] e.g. smart energy grids, syndromic biosurveillance, environmental monitoring, emergency situation awareness, digital agriculture, and smart manufacturing that are capable of producing billions of data stream from geographically distributed data sources.

Despite recent technological advances of the data-intensive computing paradigms [4] (e.g. the MapReduce paradigm, workflow technologies, stream processing engines, distributed machine learning frameworks) and datacentre clouds [5], large-scale reliable system-level software for IoT big data applications are yet to become commonplace. As new diverse IoT applications begin to emerge, there is a need for optimized techniques to distribute processing of the streaming data produced by such applications across multiple datacentres that combine multiple, independent, and geographically distributed software and hardware resources. However, the capability of existing data-intensive computing paradigms is limited in many important aspects such as: (i) they can only process data on compute and storage resources within a centralised local area network, e.g., a single cluster within a datacentre. This leads to unsatisfied Quality of Service (QoS) in terms of timeliness of decision making, resource availability, data availability, etc. as application demands increase; (ii) they do not provide mechanisms to seamlessly integrate data spread across multiple distributed heterogeneous data sources (ICOs); (iii) lack support for rapid formulation of intuitive queries over streaming data based on general purpose concepts, vocabularies and data discovery; and (iv) they do not provide any decision making support for selecting optimal data mining and machine algorithms, data application programming frameworks, and NoSQL database systems based on nature of the big data (volume, variety, and velocity). Furthermore, adoption of existing datacentre cloud platforms for hosting IoT applications is yet to be realised due to lack of techniques and software frameworks that can guarantee QoS under uncertain big data application behaviours (data arrival rate, number of data sources, decision making urgency, etc.), unpredictable datacentre resource conditions (failures, availability, malfunction, etc.) and capacity demands (bandwidth, memory, storage, and CPU cycles). It is clear that existing data intensive computing paradigms and related datacentre cloud resource provisioning techniques fall short of the IoT big data challenge or do not exist.

Hence, this special issue solicits papers related to topics including techniques for providing a secure end-to-end connection between users and data sources, QoS optimized parallel data analytic techniques, programming abstractions for extending existing data intensive computing paradigms to multiple datacentres, IoT big data application specific ontology models for capturing heterogeneous data from multiple sources, Innovative IoT big data application use cases and so on. The call for papers for this special issue received a number of submissions. After a two-phase peer review process, we have accepted four high quality papers related to the above areas of interest.

The paper titled ''Pricing and repurchasing for big data processing in multi-clouds'' by Li et. al. address the challenge of streaming big data computing service in multi cloud environments. Existing cloud pricing strategy is unfriendly for processing streaming big data with varying load. Multiple cloud environment is a potential solution however, an efficient pay-on-demand pricing strategy is demanded for processing streaming big data. They propose an intermediary framework with multiple cloud environment to provide streaming big data computing service with lower cost per load, in which a cloud service intermediary rents the cloud service from multiple cloud providers and provides streaming processing service to the users with multiple service interfaces. The also propose a pricing strategy to maximize the revenue of the multiple cloud intermediary. With extensive simulations, our pricing strategy brings higher revenue than other pricing methods.

The paper titled ''Non-intrusive anomaly detection with streaming performance metrics and logs for DevOps in public clouds: A Case Study in AWS'' by Sun et. al. address the challenges anomaly detection at user end using a non-intrusive approach. Public clouds are a style of computing platforms where scalable and elastic IT-enabled capabilities are provided as a service to external customers using Internet technologies delivering reduced costs and increased choices of technologies. Although it is possible to gain insight for the smoothness and performance of applications, it is hard to automatically detect anomalies with only data from tools like Amazon CloudWatch in complex system environments, because of the limitation of CloudWatch. Moreover, since users cannot arbitrarily access the system information in public clouds, the anomaly detection at user end has to be non-intrusive. In this paper, for tenants and DevOps practitioners in public clouds, the authors propose an anomaly detection approach, which is designed for public cloud users to deal with the case that the impacts from DevOps operations and anomalies on the metrics are same or similar. To be more specific, they report the anomaly detection on a successful public cloud, Amazon Web Service (AWS), and a representative DevOps operation, rolling upgrade. The anomaly detection technique uses Support Vector Machine (SVM) to train multiple classifiers from monitored data for different system environments, on which the log information can indicate the best suitable classifier. Moreover, the detection process aims at finding anomalies over every time interval, called window, such that the features include not only some indicative performance metrics but also the entropy and the moving average of metrics data in each window. The experimental results show that the proposed non-intrusive anomaly detection can effectively detect anomalies with the accuracy, the precision, and the recall reaching up to more than 90%.

The paper titled ''Provision of data-intensive services through energy- and QoS-aware virtual machine placement in national cloud data centers'' by Wang et. al. address the challenge of virtual machine placement across national data centres for provisioning data-intensive services such as internet of things. Many data-intensive services (e.g., planet analysis, gene analysis, etc.) are becoming increasingly reliant on national cloud data centers because of growing scientific collaboration among countries. In national cloud data centers, tens of thousands of virtual machines are assigned to physical servers to provide data-intensive services with a quality-of-service (QoS) guarantee, and consume a massive amount of energy in the process. Although many virtual machine placement schemes have been proposed to solve this problem of energy consumption, most of these assume that all the physical servers and network topologies of cloud data centres are homogeneous. However, the physical server configurations of national cloud data centers often differ significantly, which leads to varying energy consumption characteristics. In this paper, the authors explore an alternative virtual machine placement approach to minimize energy consumption during the provision of data-intensive services with a global QoS guarantee in national cloud data centers. They use an improved particle swarm optimization (PSO) algorithm to develop an optimal virtual machine placement approach involving a tradeoff between energy consumption and global QoS guarantee for data-intensive services. Experimental results based on an extended version of the CloudSim framework show that the proposed approach significantly outperforms other approaches to energy optimization and global QoS guarantee in national cloud data centers.

The paper titled ''SafeProtect: Controlled data sharing with user-defined policies in cloud-based collaborative environment'' by Thilakanathan et. al. address the challenges in privacy-aware sharing of consumer data in collaborative environments. There are many Cloud-based applications consumed by users which encourage data sharing with not only peers, but also new friends and collaborators. Data is increasingly being stored outside the confines of the data owner's machine with little knowledge to the data owner, how and where the data is being stored and used. Hence, there is a strong need for the data owner to have stronger control over their data, similar to the level of control they possess when the data is stored on their own machine. For instance, when a data owner shares a secret file with a friend, he cannot guarantee what his friend will do with the data. In this paper, the authors attempt to address this problem by monitoring and preventing unauthorised operations by the data consumer. They present a solution called SafeProtect which bundles the data owners data and policy, based on XACML, in an object. SafeProtect enforces the policies set out by the data owner by communicating with the SaaS applications to disable certain commands and/or run a background process monitor for auditability/accountability purposes. They define a protocol that enables secure data sharing in the Cloud and leverage the use of TED for authentication purposes. The authors also present a demo of the SafeProtect system by showcasing a relatively complex policy and describe how the resource is accessed by a plugin via Microsoft Word.

In summary, the papers presented in this special issue demonstrate the diversity of research in methods and techniques for processing streaming big data in datacentre clouds.

ACKNOWLEDGEMENT

We would like to thank the authors and all the reviewers for their hard work in helping us put together this special issue. We would also like to thank the Editor-in-Chief, Professor Fabrizio Lombardi and acknowledge the support of the TETC editorial staff (Alexandra Titta).

RAJIV RANJAN
School of Computing Science
Newcastle University
Tyne and Wear NE1 7RU, U.K.

LIZHE WANG
School of Computer Science
China University of Geosciences
Wuhan 430074, China

ALBERT Y. ZOMAYA
School of Information Technologies
The University of Sydney
Sydney, NSW 2006, Australia

JIE TAO
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
Karlsruhe 76131, Germany

PREM PRAKASH JAYARAMAN
School of Computer Science and Information Technology
Royal Melbourne Institute of Technology
Melbourne, VIC 3000, Australia

DIMITRIOS GEORGAKOPOULOS
School of Computer Science and Information Technology
Royal Melbourne Institute of Technology
Melbourne, VIC 3000, Australia

REFERENCES

[1] C. Perera, C. H. Liu, and S. Jayawardena, "The emerging Internet of Things marketplace from an industrial perspective: A survey," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 4, pp. 585–598, Dec. 2015.

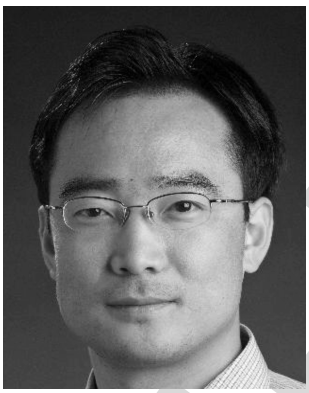
[2] L. Liu, Y. Liu, L. Wang, A. Zomaya, and S. Hu, "Economical and balanced energy usage in the smart home infrastructure: A tutorial and new results," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 4, pp. 556–570, Dec. 2015.

[3] Y. Zhou, X. Chen, A. Y. Zomaya, L. Wang, and S. Hu, "A dynamic programming algorithm for leveraging probabilistic detection of energy theft in smart home," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 4, pp. 502–513, Dec. 2015.

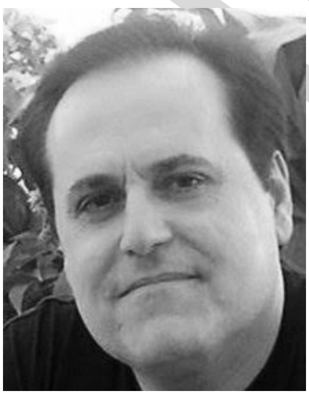
[4] L. Wang, H. Zhong, R. Ranjan, A. Zomaya, and P. Liu, "Estimating the statistical characteristics of remote sensing big data in the wavelet transform domain," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 324–337, Sep. 2014.

[5] F. Zhang, J. Cao, W. Tan, S. U. Khan, K. Li, A. Y. Zomaya, "Evolutionary scheduling of dynamic multitasking workloads for big-data analytics in elastic cloud," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 338–351, Sep. 2014.

RAJIV RANJAN was a Julius Fellow (2013-2015), Senior Research Scientist, and Project Leader with the Data61 Unit, Commonwealth Scientific and Industrial Research Organization. He is currently an Associate Professor (Reader) of Computing Science with Newcastle University, U.K. He has authored about 150 scientific papers (including over 85 journal papers) in premier venues, such as the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the ACM/IEEE World Wide Web Conference, and the IEEE TRANSACTIONS ON CLOUD COMPUTING. According to Google Scholar, his papers have received more than 5000 citations with an h-index of 26. His main research interests include autonomic management of applications in distributed systems (clouds, datacenters, and big data).

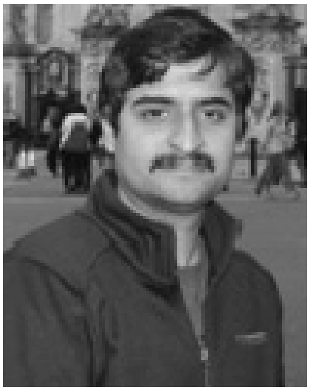


LIZHE WANG received the B.E. and M.E. (*magna cum laude*) degrees from Tsinghua University, and the D.Eng. degree from the University of Karlsruhe, Germany. He is currently a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, and a ChuTian Chair Professor with the School of Computer Science, China University of Geosciences. His main research interests include high-performance computing, e-science, and spatial data processing. He serves as an Associate Editor of the IEEE TRANSACTIONS ON COMPUTERS and the IEEE TRANSACTIONS ON CLOUD COMPUTING. He is a fellow of IET and the British Computer Society.



ALBERT Y. ZOMAYA (F'04) is the Chair Professor of High Performance Computing and Networking with the School of Information Technologies, The University of Sydney, and also serves as the Director of the Centre for Distributed and High Performance Computing. He has authored more than 500 scientific papers and articles, and co-authored and edited over 20 books. His research interests are in the areas of parallel and distributed computing and complex systems. He is the Founding Editor-in-Chief of the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING and serves as an Associate Editor for 22 leading journals. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON COMPUTERS (2011-2014). He was a recipient of the IEEE Technical Committee on Parallel Processing Outstanding Service Award (2011), the IEEE Technical Committee on Scalable Computing Medal for Excellence in Scalable Computing (2011), and the IEEE Computer Society Technical Achievement Award (2014). He is a Chartered Engineer and a fellow of AAAS and IET.

JIE TAO received the Ph.D. degree from the Munich University of Technology, Munich, Germany. In recent years, she worked intensively on grid, cloud and data-intensive computing, and the virtualization technologies. She has been a Lecturer and Research Associate with the Munich University of Technology. She is currently with the Karlsruhe Institute of Technology, Karlsruhe, Germany. She is a Principal Investigator of several research projects. She has authored or co-authored 130 articles. Her earlier research focus was mainly on parallel programming models and performance tools. She has served as a Co-Chair or a Program Committee Member of international conferences and workshops, and a Guest Editor of several journals.



PREM PRAKASH JAYARAMAN was a Post-Doctoral Research Fellow with CSIRO Digital Productivity Flagship, Australia, from 2012 to 2015. He was a Research Fellow and Lecturer with the Centre for Distributed Systems and Software Engineering, Monash University, Melbourne, Australia. He is currently a Research Fellow with the Royal Melbourne Institute of Technology University, Melbourne. His research areas of interest include Internet of Things, cloud computing, mobile computing, sensor network middleware, and semantic Internet of Things. He is one of the key contributors to the Open Source Internet of Things project (OpenIoT) that has won the prestigious Black Duck Rookie of the Year Award in 2013. He was a recipient of several awards, including hackathon challenges at the Fourth International Conference on IoT (2014) at the MIT Media Lab, Cambridge, MA, and IoT Week 2014 in London, and the best paper award at IEA/AIE-2010. He has served as a Program Committee Member of the Hawaii International Conference on System Sciences and Mobile Data Management Conferences and is a Reviewer of many distributed systems and software engineering journals, including *Future Generation Computer Systems* (Elsevier), *Concurrency and Computation: Practice and Experience* (John Wiley & Sons), the *World Wide Web Journal*, and the IEEE TRANSACTIONS ON CLOUD COMPUTING.

DIMITRIOS GEORGAKOPOULOS was a Principal Scientist with Verizon from 1990 to 1997. From 1997 to 2000, he was a Technical Manager with Microelectronics and Computer Corporation (MCC), and the Project Manager and Chief Architect in MCC's Collaboration Management Infrastructure consortial project. From 2000 to 2008, he was a Senior Scientist with Telcordia, where he helped found Telcordia's Research Centers in Austin, TX, and Poznan, Poland. Before coming to CSIRO in 2008, he held research and management positions in several industrial laboratories in the U.S. Until 2014, he was the Director of the Information Engineering Laboratory with the ICT Centre, CSIRO. He has been with the School of Computer Science and Information Technology, Royal Melbourne Institute of Technology, Melbourne, Australia, as a Professor. He is currently an Adjunct Professor with Australian National University and a CSIRO Adjunct Science Fellow. He has authored over 150 scientific papers.