Editorial for FGCS special issue: Big Data in the cloud

Victor Chang[a,*], Muthu Ramachandran[b], Gary Wills [c], Robert John Walters [c],
Chung-Sheng Li [d], Paul Watters [e]

[a] Xi'an Jiaotong Liverpool University, China
[b] Leeds Beckett University, UK
[c] University of Southampton, UK
[d] IBM, United States
[e] Massey University, New Zealand

abstract

Research associated with Big Data in the Cloud will be important topic over the next few years. The topic includes work on demonstrating architectures, applications, services, experiments and simulations in the Cloud to support the cases related to adoption of Big Data. A common approach to Big Data in the Cloud to allow better access, performance and efficiency when analysing and understanding the data is to deliver Everything as a Service. Organisations adopting Big Data this way find the boundaries between private clouds, public clouds and Internet of Things (IoT) can be very thin. Volume, variety, velocity, veracity and value are the major factors in Big Data systems but there are other challenges to be resolved.

The papers of this special issue address a variety of issues and concerns in Big Data, including: searching and processing Big Data, implementing and modelling event and workflow systems, visualisation modelling and simulation and aspects of social media.

1. Introduction

Cloud Computing and Big Data in the Cloud are becoming important topics which will warrant significant research attention in the future. The topic includes work on demonstrating architectures, applications, services, experiments and simulations in the Cloud to support the cases related to adoption of Big Data. Volume, variety, velocity, veracity and value are the major factors in Big Data systems but there are other challenges to be resolved.

For example, organisations adopting Big Data are beginning to find the boundaries between private clouds, public clouds and Internet of Things (IoT) can be very thin. The papers of this special issue address a variety of issues and concerns in Big Data, including: Adoption, Visualisation Modelling and Simulation, Cost, Security and Storage.

Papers were invited for this special issue considering aspects of this problem, including:

- Improved techniques for processing and visualisation of Big Data in the Cloud.
- Design, implementation, evaluation and services in the Cloud for Big Data.
- Systems and applications using Big Data.
- Security, privacy, trust, ownership and risk simulations for Big Data in the Cloud.
- Business and economic models, social network analyses, scientific workflows and business processes related to Big Data in the Cloud.
- Integration of Big Data in the Cloud with other technologies, including Internet of Things.

- Case studies, frameworks and user evaluations of methods for Big Data in the Cloud.
- Improved models for analysis of data storage and processing in the Cloud.

After review, a total of 11 papers have been accepted for publication in this issue.

2. Content of this issue

Of the eleven papers in this issue, four address searching and processing of Big Data, three are concerned with performing workflows and event modelling in the Cloud with four more about modelling real world systems and social media in the context of the Cloud.

The first of the four papers about data, ‘‘ActiveSort: Efficient External Sorting using Active SSDs in the MapReduce Framework’’ by Young-Sik Lee et al. [1], looks gaining an advantage in data intensive applications by using the capabilities of solid state drives (SSDs) to perform part of the work more usually carried out by the host. The paper presents an improved external sorting algorithm, ActiveSort which uses this concept of active SSDs and reports results obtained when applying this algorithm to a real active SSD platform that outperform the original Hadoop implementation by more than 35%.

The second paper, ‘‘MapReduce-based Fast Fuzzy C-means Algorithm for Large-scale Underwater Image Segmentation’’ by Xiu Li et al. [2], also considers processing of large datasets, this time considering images and videos generated by underwater cameras and other instruments. As instruments and cameras have improved in recent years, the volume of data they generate has increased rapidly. However, many researchers are only interested in unexpected events as well as plants and other lifeforms which are found underwater and one the seabed. This means they need to be able to extract the relatively small quantities of images and data which is of interest to them from the mass of data efficiently. The authors of this paper present an algorithm based on MapReduce which is able to parallelise the use of a fast fuzzy c-means (FFCM) algorithm used to identify data of interest. In addition to describing the algorithm, the authors present their evaluation of the algorithm from which they are able to claim that it represents a worthwhile improvement on traditional approaches and can handle large datasets.

The third paper, ‘‘Secure Searching on Cloud Storage Enhances by Homomorphic Indexing’’ by Shu Qin Ren et al. [3], looks at another consideration which arises in the storage of data in the Cloud. Individuals and organisations storing data in the Cloud need to be confident that their data is safe and secure. Encryption of the data a technique which appeals to them as an effective measure to prevent unauthorised access to their data. However, once the data is encrypted the Cloud service providers are unable to read it, meaning they are no longer able to supply data searching services. In this paper, a scheme is described and demonstrated as feasible which overcomes this problem by providing Cloud service providers the means to continue to provide key word searches of data which they are unable to examine direction because it has been encrypted by its owners.

The final paper on the data handling theme, ‘‘Data Adapter for Uniform Access and Transformation in NoSQL’’ by Ying-Ti Liao et al. [4], considers the consequences of the explosion of data in recent years. Large volumes of data are being stored in databases in Clouds by numerous service providers and organisations. The volume and variety of this data means that it is no longer reasonable to assume the whole of any dataset will be stored in a single database, or even in multiple instances of the same kind of database. This leads to the problem which this paper seeks to address; how to work with data which is held in a variety of locations and databases. The paper looks in particular at how to work with data which is distributed between systems where some use the relational database model and others adopt the increasingly popular ‘‘noSQL’’ approach. The paper proposes an adapter approach using which it is possible for a client to work seamlessly with both types of database.

CloudSim is a well-known open source simulation tool for researchers investigating Cloud based systems. The first paper by Wilson A Higashimo et al. [5] describes CEPSim, an extension to the CloudSim tool which permits modelling and simulation of Complex Event Processing systems in the Cloud. The paper builds on the previous work of the three authors with further consideration of CEPSim’s goals and assumptions and introducing the concept of event sets. They also present results from two experiments using CEPSim to simulate real systems.

The next paper, ‘‘A Security and cost aware scheduling algorithm for heterogeneous tasks of scientific workflow in clouds’’ by Zhongjin Li et al. [6], looks at how to operate scientific workflows securely and efficiently in the Cloud. It is often suggested that Cloud computing can make unlimited resources available to users but in reality this comes at a cost and, with service providers charging using time based pricing models, users need to pay attention to and manage their resource usage. At the same time, protecting work in progress from interference during operations in the Cloud is becoming a significant concern. Unfortunately, whilst appropriate security is generally readily available to users of Cloud services, there are associated costs. Identifying the optimal provisioning and scheduling of scientific workflows is a multi-dimensional, multiconstraint optimisation problem which has been shown to be NP-hard. In this paper, the authors propose an algorithm, SCAS, which is able to arrive at a good solution and report their experience of experiments using it the CloudSim simulation environment and three real applications.

The final paper with a Workflow theme, ‘‘Whole-exome Data Processing using Workflows on the Cloud’’ by Jacek Cala et al. [7], looks at a genomics data processing application and reports experience of porting this application from a dedicated cluster to a workflow-based system running on a public cloud. This experience is particularly useful as porting similar applications to run in Cloud environments is becoming common. In this case, the transfer from local HPC cluster to public cloud entailed reconsideration of the application to replace the former script-based HPC solution with one which is workflow-based. The paper includes some interesting observations about the outcomes of the activity, notably that the availability of VMs in the Cloud with access to fast SSD storage accounted for a significant proportion of performance gains observed and that whilst adding Cloud resources increases cost linearly, processing times reduce more slowly meaning that the fastest response times were always posted by the largest configurations, but smaller configurations were generally more cost-effective.

The special issue also includes four papers about real world applications of Cloud computing. The first looks at a system to assist with prediction of ‘‘Health Shocks’’, the second considers forecasting traffic speeds. The remaining two papers are concerned with aspects of social media.

Health Shock is the term for adverse effects on the individual affected, their family and society of serious or critical illness. Continuing improvements to healthcare monitoring and data collection together with ever reducing costs of associated hardware and infrastructure mean that the volume of health related data available is increasing rapidly. This should permit significant improvement in the prediction of health shocks. However, it seems that what is actually happening is that healthcare professionals and organisations are being overwhelmed by the mass of data available. This paper, ‘‘Cloud enabled data analytics and visualisation framework for health shocks prediction’’ by Shahid Mahmud et al. [8], describes a predictive model of Health Shock which uses a fuzzy rule summarisation technique to help with this problem. The work concentrates on rural and remote areas of Pakistan.

It is clearly desirable to be able to make accurate predictions of traffic movement speeds. However, traffic patterns vary according to the behaviour of vehicles individual roads and sections of road. It also changes through the day and from day to day. This makes predicting traffic movement rates complex and difficult. A shortcoming of most studies of traffic flow rates make their predictions is the use of samples of traffic data which means they are unable to benefit from the data which is becoming available from numerous networks. The next paper, ‘‘Monte Carlo simulation based traffic speed forecasting using historical big data’’ by Seungwoo Jeon and Bonghee Hong [9], describes an alternative traffic speed prediction technique which seeks to overcome this shortcoming and create more accurate predictions by applying simulation and statistical methods able to use all sources of information.

The final two papers of the special edition, ‘‘A personalised hashtag recommendation approach using LDA-based topic model in microblog environment’’ by Goldina Ghosh et al. [10] and ‘‘State

Transition in Communication under Social Network: An Analysis using Fuzzy Logic and Density Based Clustering Towards Big Data Paradigm'' by Fang Zhao et al. [11] are social networking related. The first examines the behaviour of users of social media and micro-blogging sites such as Twitter in particular, proposing an algorithm which uses fuzzy logic to investigate uncertainty and ambiguity in conversations. This involves observations about the nature of these conversations and how they evolve. The other considers the use of ''hashtags'', words or phrases prefixed with the ''#'' symbol inserted to attract attention, summarise content or facilitate searches. The paper proposes a personalised hashtag recommendation technique to assist users in finding content which they will find relevant or of interest.

And, finally, this editorial would not be complete without brief mention of two further papers related to this special edition which have been published as regular papers. The first of these papers, ''A model to compare Cloud and non-Cloud storage of Big Data'' by Victor Chang and Gary Wills [12] describes a model for comparisons of cloud and non-cloud storage of Big Data. The other paper, ''Cloud Computing Adoption Framework: A security framework for business Clouds'' by Victor Chang et al. [13] describes CCAF; a security framework for business clouds. Our scholarly activities have blended with our FGCS special issues. We are honoured to host invite Prof. Peter Sloot as a keynote in our IoTBD 2016 conference and as a co-host in FGCS Forum in Rome, Italy, between 23 and 25 April, 2016.

References

[1] Y.-S. Lee, L. Cavazos, S.-H. Kim, J.-S. Kim, S. Maeng, ActiveSort: Efficient external sorting using active SSDs in the MapReduce framework, Future Gener. Comput. Syst. (2016).

[2] X. Li, J. Song, F. Zhang, X. Ouyang, S.U. Khan, MapReduce-based fast fuzzy C-mans algorothm for large-scale underwater image segmentation, Future Gener. Comput. Syst. (2016).

[3] S.Q. Ren, B.H.M. Tan, S. Sundaram, T. Wang, Y. Ng, V. Chang, et al., Secure searching on cloud storage enhanced by homomorphic indexing, Future Gener. Comput. Syst. (2016).

[4] Y.-T. Liao, J. Zhou, C.-H. Lu, S.-C. Chen, C.-H. Hsu, W. Chen, et al., Data adapter for querying and transformation between SQL and NoSQL database, Future Gener. Comput. Syst. (2016).

[5] W.A. Higashimo, M.A.M. Capertz, L.F. Bittencourt, CEPSim: Modelling and simulation of complex event processing in could environments, Future Gener. Comput. Syst. (2016).

[6] Z. Li, J. Ge, H. Yang, L. Huang, H. Hu, B. Luo, A security and cost aware scheduling algorithm for Heterogeneous tasks of scientific workflow in clouds, Future Gener. Comput. Syst. (2016).

[7] J. Cala, E. Marei, Y. Xu, K. Takeda, P. Missier, Scaleable and efficient wholeexome
data processing using workflows on the cloud, Future Gener. Comput.
Syst. (2016).
[8] S. Mahmud, R. Iqbal, F. Doctor, Cloud enabled data analytics and visualization
frameowkr for health-shocks prediction, Future Gener. Comput. Syst. (2016).
[9] S. Jeon, B. Hong, Monte-Carlo simulation-based traffic speed forecasting using
historical big data, Future Gener. Comput. Syst. (2016).
[10] G. Ghosh, S. Banerjee, N. Yen, State transition in communication under social
network: An analysis using fuzzy logic and density based clustering towards
big data paradigm, Future Gener. Comput. Syst. (2016).
[11] F. Zhao, Y. Zhu, L.T. Yang, A personalized hashtag recommendation aproach
using LDA-based topic model in microblog environment, Future Gener.
Comput. Syst. (2016).
[12] V. Chang, G. Wills, A model to compare cloud and non cloud storage of big data,
Future Gener. Comput. Syst. 57 (2016) 56–76.
[13] V. Chang, Y.-H. Kuo, M. Ramachandran, Cloud computing adoption framework:
A security framework for business clouds, Future Gener. Comput. Syst. 57
(2016) 24–41.