A Technical Roadmap for Achieving Scalable Big Sensing Data Curation on Cloud

by

Chi Yang

B. Sci. (Shandong University)
M. Sci. by Research (Swinburne University of Technology)

A thesis submitted to
Faculty of Engineering and Information Technology
University of Technology Sydney

for the degree of
Doctor of Philosophy

May 2016

To my family and my friends

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has never been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written all by myself. Any help that I have received during my research work and this thesis preparation itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

Date:

Acknowledgement

First of all, I sincerely appreciate the help from my principle coordinating supervisor, Prof. Jinjun Chen, for his research suggestions, experienced supervision and continuous encouragement for my PhD study. Then, I want to express my appreciation to my co/ex-supervisors, principal research scientist Dr. Surya Nepal from CSIRO, for his supervision and generous support. Without their consistent supports and supervision, it would be very difficult complete my Ph.D study and this thesis.

I also need to express my appreciation to University of Technology Sydney (UTS) and the Faculty of Engineering and IT (FEIT) for offering me the IPRS/APA full Research Scholarship throughout my study. At the same time, appreciation should also be given to CSIRO for providing me with a top-up scholarship, supervisions and extra research facilities for my research.

My Special thanks should go to staff members, research assistants, colleagues, and friends at UTS, for their helps, suggestions, friendships and encouragements, in particular, Prof. Igor Hawryszkiewycz, Indrawati Nataatmadja, Xuyun Zhang, Deepak Puthal, Chang Liu, Ming Liu, Adrian Johannes and Nazanin Borhan.

Last but not least, I am deeply grateful to my parents Zhimin Yang and Peng Gong, my Grandma Qing Li for their long term financial support, generous understanding and emotional encouragement. Finally, I want to express my appreciation to all the friends and relatives who contribute to my life, work and experience which made my Ph.D study colourful and fruitful.

Abstract

Nowadays, big data means that data sets are so large and complex that they become difficult to process with traditional database management systems or traditional data processing tools. As important sources of big data sets, modern sensing systems generate huge volumes of sensing data beyond the ability of commonly-used software tools to capture, manage, and process within a tolerable time length. Big sensing data is prevalent in both industry and scientific research applications. The massive size, extreme complexity and high speed of big sensing data form new challenges in terms of data collection, data storage, data organization, data analysis and data publishing in real time when deploying some real world sensing systems. Cloud environment, with its massive storage, scalability and powerful computing capability, becomes an ideal platform for big sensing data processing. More and more research and industry efforts have been devoted to explore ways to process big sensing data on Cloud in order to offer better solutions for challenges brought by big sensing data. In this thesis, we will concentrate on the data curation and preparation issues under the overall theme of big sensing data processing. Especially, under the topic of big sensing data curation on Cloud, two important issues including scalable big sensing data cleaning and scalable big sensing data compression will be intensively investigated. In terms of big sensing data cleaning, a systematic approach will be developed to solve error detection and error recovery problems of big sensing data. In terms of big sensing data compression, independent techniques will be developed to reduce the size of incoming big sensing data, hence, to reduce the cost of Cloud storage, avoid big data set navigation and guarantee real time reaction. Different to previous traditional data cleaning and compression techniques, big sensing data features, the real time requirement, scalability of Cloud, will have huge influence to the techniques developed in this thesis. With those developed techniques, a detailed roadmap for achieving scalable big sensing data curation on Cloud will be proposed as our overall research outcome. Finally, the different techniques in our proposed big sensing data curation roadmap will be tested and verified with real world big sensing data sets on Cloud to show their effectiveness, efficiency and other performance gains. We aim to demonstrate that with the offered roadmap of big sensing data curation on Cloud, the typical challenges within big sensing data curation will be solved through the massive computational power and resource support from Cloud.

The Author's Publications

I have authored or co-authored 17 fully-refereed research publications during my Ph.D study, including 1 book chapter (co-authored), 6 ERA ranked A*[1] journal papers (2 as the first author), 3 ERA ranked A journal papers (1 as the first author), 5 international conference papers (1 as the first author) and other 2 high quality first author research papers under review by top research journals. The impact factor (IF)[2] of each journal paper is also associated at the end of the paper. We utilize the symbol † to indicate the first author publications which are main research outcome of this thesis.

Book Chapter:

1. Xuyun Zhang, Chang Liu, Surya Nepal, **Chi Yang** and Jinjun Chen, Privacy Preservation over Big Data in Cloud Systems. *Security, Privacy and Trust in Cloud Systems*, pages 239-258, Springer, ISBN: 978-3-642-38585-8, 2013.

Journal Articles:

2. †**Chi Yang** and Jinjun Chen, *A Scalable Data Chunk Similarity based Compression Approach for Efficient Big Sensing Data Processing on Cloud*, IEEE Transactions on Knowledge and Data Engineering (TKDE), in press, 2016. (A, IF: 2.067)

3. †**Chi Yang**, Chang Liu, Xuyun Zhang, Surya Nepal and Jinjun Chen, *A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud*, IEEE Transactions on Parallel and Distributed Systems (TPDS), vol. 26, no. 2, pp. 329-339, 2015. (A*, IF: 1.796)

[1] ERA ranking is a ranking framework for publications in Australia. Refer to http://www.arc.gov.au/era/era_2010/archive/era_journal_list.htm for detailed ranking tiers. The 2010 version is used herein. For journal papers: A* (top 5%); A (next 15%). For conference papers (no A* rank): A(top 20%).

[2] IF: Impact Factor. Refer to http://wokinfo.com/essays/impact-factor/ for details and query.

4. †**Chi Yang**, Xuyun Zhang, Changmin Zhong, Chang Liu, Jian Pei, Kotagiri Ramamohanarao, and Jinjun Chen, *A Spatiotemporal Compression based Approach for Efficient Big Data Processing on Cloud*, Journal of Computer and System Sciences (JCSS), vol. 80, no. 8, pp.1563-1583, 2014. (A*, IF : 1.106)

5. Xuyun Zhang, Wanchun Dou, Jian Pei, Surya Nepal, **Chi Yang**, Chang Liu, Jinjun Chen, *Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud*, IEEE Transactions on Computers, vol. 64, no. 8, pp.2293-2307, 2015. (A*, IF : 1.659)

6. Xuyun Zhang, Chang Liu, Surya Nepal, **Chi Yang**, Wanchun Dou and Jinjun Chen, *A Hybrid Approach for Scalable Subtree Anonymization over Big Data using MapReduce on Cloud*, Journal Computer and System Sciences (JCSS), vol. 80, no. 5, pp. 1008-1020, 2014. (A*, IF : 1.106)

7. Xuyun Zhang, Chang Liu, Surya Nepal, **Chi Yang**, Wanchun Dou and Jinjun Chen, *SaC-FRAPP: A Scalable and Cost-effective Framework for Privacy Preservation over Big Data on Cloud*, Concurrency and Computation: Practice and Experience (CCPE), vol. 25, no. 18, pp. 2561-2576, 2013. ISSN: 1532-0634. (A, IF: 0.845)

8. Chang Liu, Rajiv Ranjan, **Chi Yang**, Xuyun Zhang, Lizhe Wang, and Jinjun Chen, *MuR-DPA: Top-down Levelled Multi-replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud*, IEEE Transactions on Computers, in press, 2014. (A*, IF : 1.659)

9. Chang Liu, Jinjun Chen, Laurence T. Yang, Xuyun Zhang, **Chi Yang**, Rajiv Ranjan and Kotagiri Ramamohanarao, *Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-grained Updates*, IEEE Transactions on Parallel and Distributed Systems (TPDS), vol. 25, no. 9, pp.2234 - 2244, Sept. 2014. (A*, IF: 1.796)

10. Chang Liu, Xuyun Zhang, **Chi Yang** and Jinjun Chen, *CCBKE - Session Key Negotiation for Fast and Secure Scheduling of Scientific Applications in Cloud Computing*, Future Generation Computer Systems (FGCS), vol. 29, no. 5, pp. 1300-1308, 2013. ISSN: 0167-739X. (A, IF: 1.864)

Conference Papers:

11. †**Chi Yang**, Chang Liu, Xuyun Zhang, Surya Nepal and Jinjun Chen, *Querying Streaming XML Big Data with Multiple Filters on Cloud*, presented at the 2nd International Conference on Big Data and Engineering (BDSE 2013), pp. 1121-1127, Sydney, Australia, December, 2013.

12. Xuyun Zhang, Chang Liu, Surya Nepal, **Chi Yang**, Wanchun Dou and Jinjun Chen, *Combining Top-Down and Bottom-Up: Scalable Subtree Anonymization over Big Data using MapReduce on Cloud*, presented at the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-13), pp. 501-508, Melbourne, Australia, July, 2013. (A)

13. Xuyun Zhang, **Chi Yang**, Surya Nepal, Chang Liu, Wanchun Dou and Jinjun Chen, *A MapReduce Based Approach of Scalable Multidimensional Anonymization for Big Data Privacy Preservation on Cloud*, presented at the Third International Conference on Cloud and Green Computing (CGC 2013), pp. 105-112, Karlsurhe, Germany, October, 2013.

14. Chang Liu, Rajiv Ranjan, Xuyun Zhang, **Chi Yang**, Dimitrios Georgakopoulos and Jinjun Chen, *Public Auditing for Big Data Storage in Cloud Computing -- A Survey*, presented at the 2nd International Conference on Big Data and Engineering (BDSE 2013), pp. 1128-1135, Sydney, Australia, December, 2013.

15. Chang Liu, Xuyun Zhang, Jinjun Chen and **Chi Yang**, *An Authenticated Key Exchange Scheme for Efficient Security-Aware Scheduling of Scientific Applications in Cloud Computing*, presented at the 2011 IEEE Ninth International

Conference on Dependable, Autonomic and Secure Computing (DASC'11), pp. 372-379, Sydney, Australia, December, 2011.

Papers under Review:

16. †**Chi Yang**, Surya Nepal and Jinjun Chen, *A Scalable Non-linear Regression based Approach for Efficient Compression of Big Sensing Data on Cloud*, under review by Journal of Computer and System Sciences (JCSS), 2016. (A*, IF : 1.106)

17. †**Chi Yang** and Jinjun Chen, *A Scalable Multi-data Sources based Recursive Approximation Approach for Fast Error Recovery in Big Sensing Data on Cloud*, to be submitted, 2016.

Table of Contents

Figures

Tables