Explorations of Big Data Analytics on the NeCTAR Research Cloud and RDSI Infrastructures: Global Twittering

Prof Richard O. Sinnott

Melbourne eResearch Group,
Department of Computing and Information Systems,
University of Melbourne, Australia,
rsinnott@unimelb.edu.au

ABSTRACT

In 2013 the University of Melbourne established the "Cluster and Cloud Computing" course to train the next generation of software engineers on utilization of high-performance computing systems and development and delivery of software systems over the Cloud with specific focus on the NeCTAR Research Cloud and more recently (2014) on the combined use of the NeCTAR Research Cloud and RDSI. In this time over 120 students have been exposed to the NeCTAR Research Cloud and have undertaken a range of "big data" analytics problems. One key source of big data is social media and the courses have required a large scale programming assignment in acquiring and processing Twitter data. In 2013, 52 students were divided into teams and assigned cities across Australia. The focus was to develop Twitter harvesting applications on the NeCTAR Cloud and show how technologies such as the noSQL CouchDB solution combined with data processing algorithms such as MapReduce across the NeCTAR Research Cloud could be used to tell stories of Australian cities. In 2014, 72 students were divided into teams and assigned a global (English) speaking city (New York, London, Los Angeles, San Francisco, Washington, Boston, Chicago, Philadelphia, Dublin, Sydney, Perth, Boston, Brisbane, Toronto) and asked to compare their selected city with Melbourne. A core scenario was to explore the sentiment across the cities across a range of themes to establish whether Melbourne was/is the most livable city in the world. In undertaking this assignment students were assigned a collection of virtual machines on the NeCTAR Research Cloud (on the Melbourne node) and each assigned 250Gb of data storage on RDSI (on VicNode – www.vicnode.org.au) from a 5Tb allocation that was granted for this purpose. A key aspect of this assignment was to also show dynamic Cloud deployment whereby applications could be scaled in real time across the NeCTAR Research Cloud.

This talk will include an overview of the Cluster and Cloud Computing course and the pedagogy associated with this course; lessons learnt in using the NeCTAR Research Cloud and RDSI including what has worked and what is an area still requiring lessons to be learnt. The talk will also illustrate some of the solutions produced by the students including comparisons of cities and live big data analytics on the Cloud. These experiences (and data) are underpinning a range of other efforts within the Melbourne eResearch Group where Twitter data is used in/across eResearch projects. We will provide a brief overview of some of these topics.

ABOUT THE AUTHOR

Prof. Richard O. Sinnott is the Director of eResearch & Chair of Applied Computing Science at the University of Melbourne He was formerly technical director of the National e-Science Centre, UK; director of e-Science at the University of Glasgow; Technical Director of the National Centre for e-Social Science and Deputy Director (Technical) of the Bioinformatics Research Centre at the University of Glasgow. He has published over 200 peer-reviewed papers in conferences/journals across a wide range of computing science areas with specific focus over the last ten years in supporting communities demanding finer-grained access control (security). He teaches Cluster and Cloud Computing at the University of Melbourne.