Scaling Big Data with Nutch and Solr on Hadoop

by
Dr Dengya (Simon) Zhu
School of Information Systems,
Curtin Business School, Curtin University

DATE:	Thursday 2 October 2014
TIME:	Starting 10am
VENUE:	Curtin University, Bentley Campus, Building 402, Level 9, Room 904

Please RSVP by **COB Tuesday 30 September** to Julie Kivuyo at email Julie.Kivuyo@curtin.edu.au

For venue purposes, **RSVP is essential please**

ABSTRACT:
Nutch, Solr and Hadoop are three stable open source software packages created by the Apache Software Foundation that can be employed to deploy and implement an enterprise level search engine / information retrieval system. Nutch enables data scraping from the Internet, intranet and local file system; Solr provides enterprise level indexing and searching capability; while Hadoop is a platform built for distributed computing. The presentation will be introducing Nutch, Solr, Hadoop, and showing how to use a compiled template of Solr/Nutch to create a search engine / information retrieval system that collects data from websites of five universities of WA, and then provide the search function from the crawled data.

ABOUT THE SPEAKER:
Dr Dengya (Simon) Zhu is an adjunct prof. at the School of Information Systems, Curtin Business School, and a data miner at the Australian Taxation Office. He has a broad range of experience in government, industry, and academic research. Dr Zhu's research interests include information retrieval, data mining, machine learning, natural language processing, big data, sentiment analysis, open source software and software development. His research projects are usually practically oriented to address real world issues. He has published a number of papers in his research areas.