NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

C S S E M I N A R

Title: **An Open and Distributed Healthcare Big Data Repository for Precision Medicine**

Speaker: Dr Edward C. Cheng
Consultant and CIO of the University of Hong Kong

Date/Time: 28 April 2016, Thursday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

Chaired by: Dr Tan Kian Lee, Shaw Senior Professor, School of Computing (tankl@comp.nus.edu.sg)

Abstract:

A large amount of data is collected from patients in every hospital every day. Most of the large hospitals now have an integrated Clinical Data Repository (CDR) that hosts patient data including patient demographics, clinical consultation notes, test results, imaging reports, prescriptions, diagnoses, treatment methods and outcomes. The insights hidden in this wealth of data can be the key to improve patient care and advance healthcare research. This is information from one hospital, what if we can combine the medical records into a gigantic library including CDRs from multiple hospitals? The ability to draw knowledge from such a resource would enable us to achieve the goal of Precision Medicine, where effective treatment methods and prescription are derived by focusing on individual patient's unique profile of history, life-style and genome. Research has been launched to create this kind of Big Data store by uploading information from various CDRs into a common data repository. The problem with this approach is two-folded, namely the complexity and the sheer volume of data. First, the data definition of the CDR from different hospitals varies. This creates a challenge of combining data from different sources with different data definitions. The result is a huge data dump with major obstacles when trying to piece together the data in a meaningful way and to perform data mining. Secondly, everyday there are over 20-50 GB of data generated from each large hospital, to upload data streams even from tens of hospitals to a centralized database would require expensive high-speed connections and a challenging indexing method to reference the data, let alone if we are dealing with hundreds of hospitals.

In this paper, an open and distributed Big Data model is introduced. By applying Data Unification (DU) technology, instead of physically bringing data from multiple CDRs, we connect a Healthcare Big Data (HBD) Engine to the numerous data sources and put a DU adapter in front of each CDR. The DU adapter wraps around the CDR and maps its attributes and metadata definition to those that are known to the HBD Engine. When a user queries

this Healthcare Big Data store, the HBD Engine will disintegrate the query and distribute it to all of its data sources for parallel processing. Through the DU Adapter, the query to each data source is mapped to the native CDR for retrieval, and the retrieved result is mapped back to the data definition of the HBD and returned to the HBD Engine. The HBD Engine is responsible for integrating the results back to the user. This distributed, parallel-processing model will guarantee a satisfactory performance and scalability of hundreds or thousands of CDRs in the HBD. It also removes the security threat and risk of data breach when all data is managed by one centralized Big Data center.

Biodata:

Edward C. Cheng is a consultant and CIO of the University of Hong Kong, assigned to be CIO and GMIT of the HKU-Shenzhen Hospital. Before joining HKU, Edward was the President & CEO of EGI Technologies, Inc., a CA-based technology company which provides a cloud-computing and Big Data solution for global institutions. Prior to that Edward was an R&D Director at Oracle Corporation in the US. He also held senior management positions at Hewlett Packard Co. and Digital Equipment Corporation.

His research interests include Big Data, Cloud-computing, medical information systems, process automation, collaborative management, social enterprise, high performance systems, journaling and recovery, and distributive databases. He has published numerous papers and authored three worldwide patents in the above areas. Edward was on the Stanford University Database Research team for three years. Edward received his PhD in CS from the University of London. He has an MBA from the University of California, Berkeley, and an MBA from Columbia University, New York.

Edward is the lead inventor of the LSM-Tree, which is implemented in all major Big Data Management System today including Google, Facebook, Oracle, Microsoft, Hadoop, Cassandra and others.