NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Supporting Efficient Database Processing in MapReduce**

Speaker: Mr Lu Peng

Date/Time: 19 November 2014, Wednesday, 03:00 PM to 04:30 PM

Venue: MR1, COM1-03-19

Supervisor : Dr Ooi Beng Chin, Professor, School of Computing

Abstract:

Cloud computing has emerged as a multi-billion dollar industry and as a successful paradigm for web-scale application deployment. Represented by the MapReduce processing model, MPP (Massively Parallel Processing) systems form a critical component of the cloud software stack. Hailed for its high scalability, massive parallelism, and effectively programmable interface, the MapReduce paradigm is widely recognized as a revolutionary advancement in large scale computation. However, due to the heterogeneity and massiveness nature of data in the Cloud, current Cloud systems trade rigorous data management functionalities for better versatility and scalability. On one hand, the absence of comprehensive data model and access methods, which have been developed extensively for relational database management systems (RDBMSs), has affected MapReduce-based system?s applicability to a wider variety of real world analytical tasks. On the other hand, due to the complexity of processing logic layers in its system architecture, RDBMSs fail to provide desirable scalability and elasticity.

The overarching goal of this dissertation is to exploit the opportunity for a better marriage of RDBMS technologies and Cloud Computing systems. This dissertation shows that with careful choice of design and features, it is possible to architect a large scale system that syncretizes the ef cient access methods of RDBMS and the powerful parallelized processing of MapReduce. This dissertation advances the research in this topic by improving two critical facets of large scale data processing systems. First, we propose an architecture to support the usage of DBMS-like indexes in MapReduce systems to facilitate the storage and processing of structured data. We start with devising a bitmap-based indexing scheme that provides superior space ef ciency, and improves the performance of MapReduce programs on a speci c category of data. We then generalize the index ap- plication, and propose a generalized index framework for MapReduce systems to handle large data and applications. Second, we propose models and techniques to incorporate the power of MapReduce with parallel database system technologies in query processing.