NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Write-Intensive Data Management In Log-Structured Storage**

Speaker: Mr Wang Sheng

Date/Time: 28 September 2016, Wednesday, 01:30 PM to 03:00 PM

Venue: Executive Classroom, COM2-04-02

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

Abstract:

As the rapid development of information technologies, huge amounts of data are generated every day. Real-world workloads are becoming write-intensive and large-scale. On one hand, world-wide applications have large user bases acting simultaneously. On the other hand, large and cheap storage drives allow us to capture high-volume data, e.g., user activity logs and sensor readings. This write-heavy tend poses new challenges to data management solutions, where databases are required to provide high throughput for write operations while preserving read performance.

In this thesis, we work towards designing solutions for managing write-intensive workloads with the adoption of log-structured techniques. We first propose a distributed log-structured storage, providing high write-throughput for key-value operations. It removes write bottleneck by unifying data and log repositories, and supports fast failure recovery. Second, we design a novel indexing method on top of log storage to support efficient range queries. This method works well for observational data, which is a common and important type of write-intensive source. Instead of creating induced clustering for all records in typical indexing methods, it directly utilizes intrinsic clustering property in original data source, and relies on sequential access to ensure query efficiency. After that, we provide an extended solution for indexing multi-dimensional observational data. It overcomes the data sparseness in multi-dimensional spaces, and minimizes space ?over-coverage? introduced by conventional spatial indexing methods.

Extensive experiments have been conducted using both real-world and benchmark workloads. First, we observe that log-structured storage is a amendable choice for write-intensive workloads, in which write throughput is of great importance. Second, we confirm the effectiveness of indexes exploiting intrinsic clustering property inside original data sources. Third, we show that even our approaches are optimized for write throughput they still preserve good read efficiency.