Tutorial: Deep Dive into Apache Cassandra: Theory, Design, and Application

Lewis Tseng, Haochen Pan, Yingjian Wu
Boston College
Boston, USA
{lewis.tseng, haochen.pan, wuit}@bc.edu

Tutorial Abstract

Distributed storage systems are fundamental to many Internet-scale applications. To achieve high performance, modern storage systems often choose to provide only weaker guarantees (e.g., weak consistency and eventual correctness). Moreover, to deal with semi- or unstructured and unpredictable data, these systems mainly support a flexible data scheme and simple interface. As a result, it requires a new way of thinking about storage and data modeling.

This tutorial consists of a series of hands-on and interactive exercises for beginners that have minimum or even no experience in distributed storage systems. We will first give an overview of distributed storages and new storage paradigms like NoSQL/NewSQL. Then we discuss their key properties (e.g., consistency, availability, fault-tolerance, etc.) and an important trade-off (i.e., CAP theorem). Finally, we will walk through exercises of installing and configuring Cassandra. If time permits, we will build a simple application on top of Cassandra.

Tutorial Outline

The tutorial will follow the outline below:

- Brief introduction to distributed storage systems
- Brief introduction to NoSQL/NewSQL and their properties
- Brief introduction to CAP theorem
- Exercises: install, configure, and develop applications on top of Cassandra

Audience

The tutorial will be accessible to anyone with a background of basic knowledge on algorithms and programming. We do *not* assume any background in cloud computing or distributed systems. Some knowledge of Linux would help.

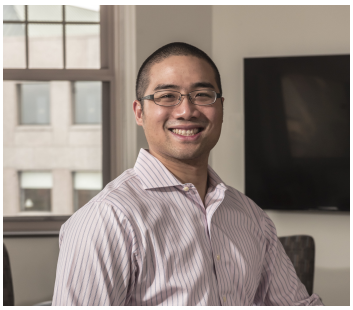
We welcome everyone to attend, but the tutorial is mainly designed for people who are interested in large-scale distributed storage systems and data-intensive storages that require ultra-high availability and low latency.

Learning Goals

After the tutorial, you are expected to learn:

- What is NoSQL/NewSQL? What is Cassandra?
- What are the trade-offs of large-scale distributed storage systems?
- How do I install, configure, and develop applications on top of Cassandra?

Bio



Lewis Tseng is currently an assistant professor in Computer Science department at Boston College. Before that, he spent a year and a half as a researcher at Toyota InfoTechnology Center. He received a B.S. and a Ph.D. degree both in Computer Science from the University of Illinois at Urbana-Champaign (UIUC) in 2010 and 2016, respectively. His research broadly lies in the intersection of fault-tolerant computing and distributed computing. He won the best paper award in the International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS) 2017.

Haochen (Roger) Pan is a junior at Boston College pursuing his dual degree in Computer Science and Mathematics. His research interests include distributed computing & systems, Blockchain-based systems, and vehicular ad-hoc networks. He received the Sophomore Scholar Award, the Advanced Study Grant, and the Undergraduate Research Fellowship from Boston College.

Yingjian (Steven) Wu is a senior at Boston College pursuing his dual degree in Computer Science and Mathematics. His research interests include distributed computing & systems, Blockchain-based systems, and Big Data infrastructure internal design. He received IEEE Comsoc Student Travel Grant for Globecom 2019. He received the Advanced Study Grant for thesis research and Undergraduate Research Fellowship from Boston College.