

A complex network diagram with nodes and edges. Nodes are represented by circles in dark blue, red, and grey. Edges are thin lines connecting the nodes, with red lines forming a dense web and grey lines forming a more sparse structure. The background is a light blue-grey gradient.

# **BIG DATA MANAGEMENT**

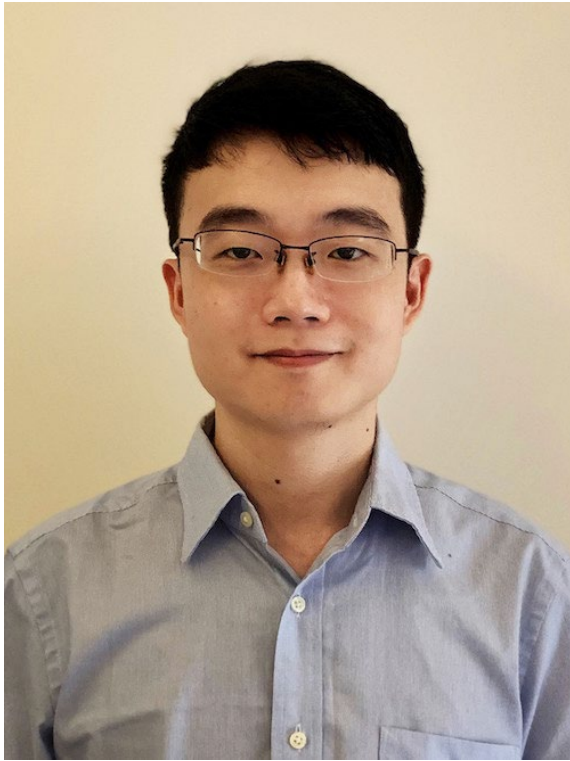
CZ/CE4123

# Course Overview



# COURSE INSTRUCTOR

LUO Siqiang (Assistant Professor at SCSE)



## Email

[siqiang.luo@ntu.edu.sg](mailto:siqiang.luo@ntu.edu.sg)

## Research area

Big data / data management

## Office

N4 Level 2 c-110

Teaching Assistant to help the **course project**

Wang Fan  
FAN008@e.ntu.edu.sg



For **all the lecture related questions**, please directly email me at [siqiang.luo@ntu.edu.sg](mailto:siqiang.luo@ntu.edu.sg)

For **project related problems**, you can consult TAs for specifics.

# WHAT IS BIG DATA MANAGEMENT?



# WHAT IS THIS COURSE ABOUT?



**Understand  
important concepts  
of big data**



**Analyze important  
big data processing  
techniques**



**Explain various big  
data systems**

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ☐ Understand what is big data

Solving math?

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ☐ Understand what is big data

Solving math?

No!



# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ☐ Understand what is big data

***Big data 5V's***

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ Consider a scenario:
  - ❑ You are a big data engineer/ analyst in Amazon, your boss assigns you a task:



Hey, can you write some code to help analyze the best-seller in this season?



**Sure! I can scan each record and get the selling frequency of each product, and then get the item with the max frequency!**

# AFTER THIS COURSE, YOU WILL BE ABLE TO...



Umm ... Not bad, but forgot to tell you.  
We have 1 trillion sale-item records...  
Will your method be efficient?

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ You will learn some solutions from the course



**MapReduce!**

- ❑ Use a few lines of code to efficiently analyze the “best seller” in retail applications in a distributed system!

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ Consider another scenario:
  - ❑ The other day, you receive an urgent task from your boss:



Hey, can you do me a favor to sort the one trillion record based on a specific attribute (e.g., product ID or user ID)?



Easy! I get quite familiar with different kinds of sorting algorithms such as quick-sort, heap-sort, ...

# AFTER THIS COURSE, YOU WILL BE ABLE TO...



Umm... I am afraid we do not have a machine that can hold all the data in main memory...



## AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ After the course, you will probably have some solutions



**External Sorting**

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ Next scenario:
  - ❑ 2 years later, you are promoted to a tech leader. Your boss wants you to design a system for easy queries of product item information.



Can you design a system to hold one trillion records, so that user can easily query the information of any given product?



**Yeah! I get familiar with SQLServer, ...**

# AFTER THIS COURSE, YOU WILL BE ABLE TO...



Umm... Can we do it more scalable?

## AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ After the course, you will probably have alternative solutions



**NoSQL Key-Value Stores**

# AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ Next scenario:
  - ❑ 2 years later, your boss wants you to redesign the system.



Hey, can you redesign the system to hold the product-purchase data, so that user can easily **query and filter** the product information based on ID and name?



**Well, this time maybe SQLServer is a better solution?**

# AFTER THIS COURSE, YOU WILL BE ABLE TO...



I do not disagree. But can we do even better given that the product may have hundreds of properties?



## AFTER THIS COURSE, YOU WILL BE ABLE TO...

- ❑ After the course, you will probably have alternative solutions



**Column Store**

# PREREQUISITES

Course CZ2007:

Introduction to Databases

~~CE/CZ4031:~~

~~Database system principles~~

# THIS COURSE IS

- ❑ NOT a programming language course
  - ✓ Will not teach SQL (had learnt it in CZ2007)
  - ✓ Will not teach C or Java (may have learnt it in other courses)
  - ✓ Though we assume you understand **one of** basic SQL or C or Java
  
- ❑ NOT a traditional database course
  - ✓ Will not teach relational database (may recap if necessary)

# BIG DATA MANAGEMENT – COURSE OVERVIEW

We will discuss interesting big data techniques!



Big Data  
5V's

Memory  
Hierarchy

Column  
Store

Distributed  
MapReduc  
e Systems

NoSQL  
Key-Value  
store

# BIG DATA MANAGEMENT – COURSE OVERVIEW

Most of the techniques are cutting-edge techniques in big data!

No text books – Slides contain everything



Big Data  
5V's

Memory  
Hierarchy

Column  
Store

Distributed  
MapReduc  
e Systems

NoSQL  
Key-Value  
store

# LECTURE STYLE

- ☐ The purpose of the course is to expand your vision, both conceptually and technically.
- ☐ I tend to encourage questions and (open-ended) discussions during the class.
- ☐ I tend to link the techniques to some real industrial systems.

# TUTORIAL STYLE

In the tutorial of this course, you will be

- ❑ Taught with solving some theory questions related to big data techniques;
- ❑ Hands-on tutorials to walk you through some widely adopted big-data systems, such as Hadoop (used by many companies) and RocksDB (used by Facebook/Meta). You will not be examined on the procedure; instead, our purpose is to give you some resources which might be useful when you join the industry in the future.

# COURSE SCHEDULE

**Week 1 – Week 13:** Course lectures

Venue: LT27

Time: 15:30pm – 17:20pm on Tuesday

**Week 3 – Week 13:** Tutorials

Venue: LT27

Time: 11:30am – 12:20pm on Monday

**Quiz:** Week 10 tutorial



# EVALUATION (TENTATIVE)

**1 Quiz:**

25%

**1 Group Project:**

25%

**Final:**

50%

# QUESTIONS WANTED!!

I welcome all kinds of questions during the course!!

Questions wanted!

**Let's start the journey!**