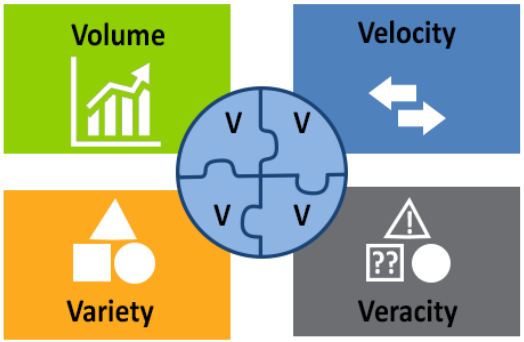
Lecture slides of the course Introduction to Big Data

COURSE INTRODUCTION

Le Ngoc Thanh – Nguyen Ngoc Thao

{lnthanh, nnthao}@fit.hcmus.edu.vn

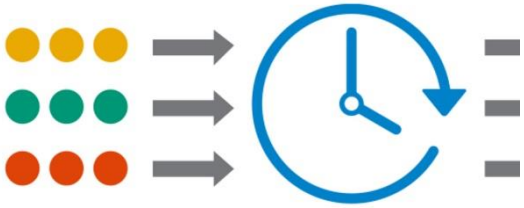
Main topics in this course



Definitions, characteristics, and case studies, etc.

Hadoop DFS and MapReduce

Spark modules: Spark SQL, MLlib, GraphFrames, etc.



Stream processing

NoSQL databases and MongoDB

Weekly schedule

Week	Lecture topics
1	The Fundamentals of Big Data - Basic concepts and Characteristics - Frameworks and Technologies
2-3	Hadoop Fundamentals: Basic concepts and Ecosystem Hadoop Distributed Filesystem: HDFS daemons I/O operations
4	Hadoop MapReduce: MapReduce daemons Practice with Apache Hadoop
5	Spark Fundamentals: Basic concepts and Ecosystem Spark APIs
6	Advanced Analytics with Spark Practice with Apache Spark
7	**Midterm examination (Theory part)**
8	NoSQL databases: MongoDB in conjunction with Spark
9	Stream processing: Spark Streaming
10-12	Seminar

Assignments, Exams, and Grading

- Grading

Midterm	Final	Seminar	Labs
25%	25%	30%	30%

- Midterm exam topics
  - Big Data fundamental, Apache Hadoop, Apache Spark
- Final exam topics
  - Spark APIs, NoSQL databases, Spark Streaming, and projects presented in the seminar
- Paper-based part (Mid: 60 mins – Final: 90 mins)
  - In-class and open-book with only 2 A4 sheets.
- Coding part (All: 60 mins)
  - At-home and open-book. Frameworks prepared in advance.

Assignments, Exams, and Grading

- Grading

Midterm	Final	Seminar	Labs
25%	25%	30%	30%

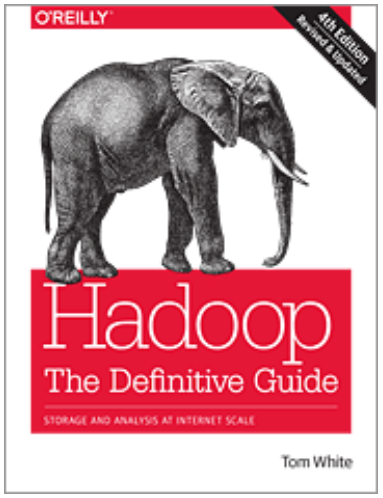
- Group assignments: 04 members/group recommended
- Seminar
  - Research on Big data topics and give oral presentation
- Lab work on Hadoop, Spark, and MongoDB
  - Install and explore the tools with simple case studies

Course regulations

- For any kind of cheating and plagiarism, students will be graded 0 for the course. The incident is then submitted to the school and university for further review.
- Students are encouraged to form study groups to discuss on the topics. However, individual work must be done and submitted on your own.
- Students absent for the mid-term or final exam are considered unqualified for course completion.
- Students must accumulate at least 10% of course credits for lab work.
- [This is for class 21KHMT1] Students will have several before-class assignments during the course. There will be a 1% penalty for each assignment missed.

List of reference books

- Tom White. 2015. *Hadoop: The Definitive Guide.* O'Reilly Media, Inc.



- Bill Chambers and Matei Zaharia, 2018. *Spark: The Definitive Guide: Big Data Processing Made Simple*. O'Reilly Media, Inc.