IT5427E
Big Data Integration and Processing

HomePage

Class Information

Time: 07:30-11:30, Monday & Thursday
Location: B1-504

Instructor: Vũ Tuyết Trinh
School of Information Technology and Communication
Hanoi University of Science and Technology
Email: trinhvt@soict.hust.edu.vn

Description

The course aims to provide students with knowledge related to the integration and processing of big data. The goal of data integration is to provide users with unified access to data published in a variety of data sources. Meanwhile, big data processing refers to the manipulation of input data. In the context of big data, these tasks face the challenges of volume, velocity, variety, and veracity, which require the use of new technologies and services to integrate and process data. To achieve the mentioned goals, the course is designed into two main parts. First, students are introduced to techniques for integrating big data such as schema matching, record linkage, and data fusion. Next, students are introduced to the Apache Spark for processing big data, with components such as streaming data processing, data query, machine learning library.

Grading

- MidleTerm 40%
- Final exam 60%

Text and Reading

1. Xin Luna Dong, Divesh Srivastava (2015). Big Data Integration. Morgan & Claypool publishers
2. AnHai Doan, Alon Halevy, Zachary Ives (2012). Principles of Data Integration. Morgan Kaufmann-Elsevier
3. Rick Sherman (2015). Business Intelligence Guidebook – From Data Integration to Analytics. Morgan Kaufmann-Elsevier
4. Matei Zaharia, Bill Chambers (2018). Spark: The Definitive Guide: Big Data Processing Made Simple. O'Reilly Media

Useful website/resources

- will be provided during the class.

Tentative Plan

Week	Topics	Materials
8/7	Data Integration Big Data Integration	1_DataIntegration.pdf
11/7	Describing Data Sources	2_SourceDescription.pdf
	Adaptive Query Processing	3_QueryProcessing.pdf
15/7	No Class	
18/7	Data: integrity, security, privacy, .. gouvernance	2_DataGouvernance.pdf
22/7	Adaptive data integration and processing	3_Adaptive.pdf
	Big Data Integration & Processing (TBC)	
25/7	Presentation* Topics: industry-specific, technology/technique, R and/or D focus	
29/7		
	Test**	

*Presentation individual/group

- group 1-3 students: registration on July 10, 2024: Project.xlsx
- Discussion on topics & plan: now-July 18, 2024: group channels
- Presentation: July 20-31, 2024
- Final report: August 15, 2024

** Test: after the last presentation (date - TBD)