國立東華大學

教學計劃表　Syllabus

課程名稱(中文) Course Name in Chinese	大數據系統			學年/學期 Academic Year/Semester	112/1
課程名稱(英文) Course Name in English	Big Data Systems				
科目代碼 Course Code	CSIEM0410	系級 Department & Year	碩士	開課單位 Course-Offering Department	資訊工程學系
修別 Type	選修 Elective	學分數/時間 Credit(s)/Hour(s)	3.0/3.0		
授課教師 Instructor	/吳秀陽				
先修課程 Prerequisite					
課程描述　Course Description					

The term "big data" is now commonly used to mean that the growth of data in volume, velocity, variety and veracity are in such an unprecedented scale that traditional database management systems can no longer handle it properly. New technologies, artificial intelligence(AI), machine learning(ML) and internet of things(IoT) in particular, rely heavily on the processing of huge data sets. Online services (ChatGPT, YouTube, Meta, IG, …) need to handle hundreds of millions of users issuing billions of request at the same time. We therefore need new technologies (big data processing/analysis) and new tools (big data systems) to deal with extremely large data sets and service requests. This is an introductory course on big data concepts, processing, analytics and systems. You will learn the latest development in big data technologies and get hands on experience in using popular open source big data systems such as Hadoop, Spark, HBase, MongoDB, Neo4j, Kafka, Flink, etc. The objectives of this course can be summarized as follows.

. Understand big data concepts, challenges and trends.
. Learn the technological foundations of big data science and engineering.
. Learn the principles and practices behind popular open source big data systems.
. Get hands on experiences of using open source big data systems for solving big data problems.

This is a lecture-oriented course. The system part of the course will be executed through in-class example discussion, homework assignments and term project. Due to the time limit, the lectures will focus mostly on the technological innovation of each system rather than how to use them. With brief introduction to the basic operations of various big data systems, students are expected to learn to use them on their own.

Regular Topics

The topics to be covered in the lecture are listed as follows (**: topics to be covered depending on the time and progress):

. Introduction
. General purpose big data platforms
. Big data storage architecture and systems
. Big data systems for structured/semi-structured data
. Big graph processing
. RDF processing systems**
. Big stream processing
. Big data pipelining tools**
. Big data ETL(extract, transform, and load) tools
. Big data analytics, other systems and trends**
. Open data**
. Big data system landscape**

Special Topic(s)

Based on current practices and emerging trends, we will select one or two special topics to provide a brief overview (if time permits, of course). This semester, the special topic we plan to talk about is big data and AI (Artificial Intelligence):

. Relationship between big data and AIML(Artificial Intelligence and Machine Learning)
. How ChatGPT and other similar Large Language Models(LLMs) work?
. LLaMA, the open source LLM from Meta
. The future of big data and generative AI

課程目標　Course Objectives

大數據處理在當前資訊爆炸的時代，已成為必備技能。特別是席捲全世界的人工智慧與機器學習浪潮，都必須仰賴對大數據的解析、分類、辨識和歸納。本課程探討大數據特性，讓同學了解大數據處理尖端技術發展脈絡，深入探索各種大數據系統和工具的運作原理和應用實例，培育同學們具備未來將大數據處理技術應用在任何領域所需的理論知識和實務技能。

In the age of information explosion, big data processing is already a must-have skill. Especially on the new waves of AI and machine learning, all systems rely heavily on big data analysis, classification, recognition and induction. The purposes of this course are to study big data characteristics, to understand the evolution of big data processing technologies, and to explore the underlying principles and real-world applications of big data systems/tools. Students will learn the theoretical knowledge and practical skills necessary for applying big data technologies on any future application domains.

系專業能力 Basic Learning Outcomes		課程目標與系專業能力相關性 Correlation between Course Objectives and Dept.’s Education Objectives
A	統合資工知識技術之能力Ability to integrate knowledge and technologies of computer science and information engineering.	●
B	設計技術理論驗證實驗之能力Ability to design and conduct science experiments and to validate hypotheses.	●
C	資訊軟硬體設計開發之能力Ability to design and develop computer software and hardware.	●
D	團隊專案開發之能力Ability to design and develop team projects.	○
E	批判性思考與創新研發之能力。Ability of analytical thinking, creative research planning, and innovative development.	●

圖示說明Illustration ：● 高度相關 Highly correlated ○中度相關 Moderately correlated

授 課 進 度 表 Teaching Schedule & Content

週次Week	內容 Subject/Topics	備註Remarks
1	Course Introduction: . Course description . Objectives . Syllabus . Textbook and references . Assignments . Independent study . Exam . Term project	
2	Introduction . Data -> Knowledge -> Intelligence . What is a Big Data? . Why Big Data? . Examples of Big Data . The opportunities and challenges for Big Data	
3	General purpose big data platforms I . Distributed and cluster computing . Apache Hadoop . Cloudera(CDH, Cloudera Distribution for Hadoop) . MapReduce and algorithm design	
4	General purpose big data platforms II . Apache Spark and in-memory computation . High Performance Computing ClusterÂ (HPCC), also referred to as DAS(DataÂ AnalyticsÂ Supercomputer)	
5	Big data storage architecture . Distributed nodes . Scale-out NAS . All-solid-satae drive (SSD) arrays . Object-based storage . DNA storage	
6	Big data storage systems . Distributed file systems and big data storage . Google GFS and Apache HDFS . Cloud Storage . Data lake** . Big data storage security**	

7	Big data systems for structured/semi-structured data I . SQL, NoSQL, NewSQL, Distributed SQL . Apache HBase, Cassandra, CouchDB, Drill, Impala, Hive Spark SQL, DataFrames, Datasets MongoDB Google BigQuery, Spanner, F1 Presto	
8	Big data systems for structured/semi-structured data II . Spark SQL, DataFrames, Datasets . MongoDB . Google BigQuery, Spanner, F1 . Presto	
9	期中考試週 Midterm Exam No midterm. Term project proposal and selected tool demo instead.	
10	Big graph processing I . The challenges of big graph processing . Pregel family of systems(BSP, Pregel, Giraph) . GraphLab family of systems(GraphLab, PowerGraph, GraphChi)	
11	Big graph processing II . Spark GraphX, GraphFrames . Neo4j graph database . Titan distributed graph database RDF processing systems** . NoSQL-based RDF systems . Hadoop-based RDF systems . Spark-based RDF systems	
12	Big stream processing I . The challenges of big data streaming . Spark Streaming, Structured Streaming	
13	Big stream processing II . Apache Storm, Samza, Flink . Apache SAMOA . Big data streaming applications	
14	Big data ETL(extract, transform, and load) tools . Apache Airflow . Apache Kafka Big data pipelining tools** . Pig Latin . Tez	
15	Big data analytics, other systems and trends . Google Cloud Platform (GCP) vs Amazon Web Services (AWS) . RapidMiner . KNIME . Tableau . R Language and RStudio . Open data	
16	Special topic: Big data and AI . The relationship between big data and AI . How do big data and AI work together? . Powerful synergy of big data and AI . How ChatGPT and other similar Large Language Models(LLMs) work? . LLaMA, the open source LLM from Meta . The future of big data and generative AI	
17	Student presentation	
18	期末考試週 Final Exam	

教 學 策 略 Teaching Strategies

☑ 課堂講授 Lecture ☐ 分組討論Group Discussion ☐ 參觀實習 Field Trip

☑ 其他Miscellaneous: Students will get hands-on experience with popular open source big data

教 學 創 新 自 評 Teaching Self-Evaluation

創新教學(Innovative Teaching)

☐ 問題導向學習(PBL) ☐ 團體合作學習(TBL) ☐ 解決導向學習(SBL)

☐ 翻轉教室 Flipped Classroom ☐ 磨課師 Moocs

社會責任(Social Responsibility)

☐ 在地實踐Community Practice ☐ 產學合作 Industy-Academia Cooperation

跨域合作(Transdisciplinary Projects)

☐ 跨界教學Transdisciplinary Teaching ☐ 跨院系教學Inter-collegiate Teaching

☐ 業師合授 Courses Co-taught with Industry Practitioners

其它 other: Play with open-source big data tools on VM.

學期成績計算及多元評量方式 Grading & Assessments									
配分項目 Items	配分比例 Percentage	多元評量方式 Assessments							
		測驗會考	實作觀察	口頭發表	專題研究	創作展演	卷宗評量	證照檢定	其他
平時成績 General Performance	15%			✓	✓				Independent study and presentation
期中考成績 Midterm Exam	0%								
期末考成績 Final Exam	25%	✓							
作業成績 Homework and/or Assignments	35%		✓						
其他 Miscellaneous (Term Project)	25%		✓		✓	✓			

評量方式補充說明
Grading & Assessments Supplemental instructions

Independent Study and Presentation
. All students are to conduct independent study on self-selected topics (discussed with me first).
. Pick an open source big data system not discussed in the class as the study target.
. Prepare a presentation and a demonstration of the system in class.
. Every student must present and demo.

Term Project
. There will be a modest scale term project for you to show your creativity.
. May use any big data tools for your project.
. Prepare a project proposal and a tool(s) demo on VMs by the end of the midterm week.
. Prepare a project demo at the end of the semester to explain your project to me.
. Turn in the project and report one week after the final exam.

教科書與參考書目（書名、作者、書局、代理商、說明）
Textbook & Other References (Title, Author, Publisher, Agents, Remarks, etc.)

No required textbook.

References:
. Jawwad Ahmad Shamsi and Muhammad Khojaye. Big Data Systems: A 360-degree Approach. Chapman & Hall/CRC, 2021.
. Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry, Amir H. Gandomi. Big Data: Concepts, Technology, and Architecture. Wiley, 2021.
. Sherif Sakr. Big Data 2.0 Processing Systems: A Systems Overview, 2nd Edition, Springer, 2020.
. S. Sasikala and D. Renuka Devi (Authors), Raghvendra Kumar (Editor). Research Practitioner's Handbook on Big Data Analytics. Apple Academic Press, 2023.
. Guido Dartmann, Houbing Song, et al. Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things. Elsevier Science Publishing Co Inc, 2019.
. Kai Hwang and Min Chen. Big Data Analytics for Cloud, IoT and Cognitive Computing. John Wiley & Sons Ltd., 2017.
. Tom White. Hadoop: The Definitive Guide, 4th Edition, O'reilly, 2015.
. Jure Leskovec, Anand Rajaraman, Jeff Ullman. Mining of Massive Datasets. Cambridge University Press, 2010~2014.
. Mohammed J. Zaki and Wagner Meira JR. Data Mining and Machine Learning - Fundamental Concepts and Algorithms, 2nd Edition. Cambridge University Press, 2020.

課程教材網址(含線上教學資訊,教師個人網址請列位於本校內之網址)
Teaching Aids & Teacher's Website(Including online teaching information. Personal website can be listed here.)

Course homepage: http://web.csie.ndhu.edu.tw/showyang/BigDataSys2023f/index.html
Instructor's homepage: http://web.csie.ndhu.edu.tw/showyang/index.html

All lecture notes will be available online.

其他補充說明 (Supplemental instructions)