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	113/1
課程名稱(英文) Course Name in English	Big Data Systems				
科目代碼 Course Code	AIIA50050	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

根據Statista的最新估計，每天有402.74百萬TB(terabytes)資料產生，今年將有147ZB(zettabytes)，2025年更將高達181ZB。這個資料累積速度是2010年2ZB的74倍！而全世界所有資料總和的90%是過去兩年間產生。這樣高速產生、大量累積的資料，正是本課程所謂的大數據(big data)。主要原因是愈來越方便多元的網路服務，社群媒體(social media)的盛行讓多媒體資料全球氾濫，近年來物聯網(IoT)和人工智慧(AI)的廣泛應用，更是將大數據的產生和需求，推向前所未有的高峰。面對這樣的局面，傳統資料庫和資料倉儲技術已經不敷使用。大數據系統和技術(big data systems & technologies)應運而生，特別是配合大數據分析(big data analytics)，能夠大幅提升組織和企業競爭力。

本課程介紹大數據概念，運作架構、和相關系統工具。所介紹系統以開源軟體為主，內容包括
．通用大數據處理平台(General purpose big data platforms): Hadoop/MatReduce, Apache Spark, HPCC, ...
．大數據儲存架構和系統(Big data storage architecture and systems): distributed storage and file system, Google GFS, Apache HDFS, GlusterFS, ...
．結構與半結構大數據處理系統(Big data systems for structured/semi-structured data): NoSQL/NewSQL/Distributed SQL, Apache HBase, Cassandra, Hive, MongoDB, ...
．圖形大數據處理系統(Systems for big graph processing): BSP, Pregel, Giraph, Spark GraphX, Neo4j, Titan, ...
．串流大數據處理系統(Systems for streaming big data): Spark Streaming, Structured Streaming, Apache Storm, Samza, Flink, SAMOA, ...
．大數據管道、ETL、工作流、和協作系統(Big data pipelineing, ETL, workflow and orchestration systems): Apache Airflow, Apache Kafka, Apache NiFi, Dagster, Prefect, ...
．大數據分析與進階議題(Big data analytics and advanced topics)

同學們將利用虛擬機管理工具(Virtual Machine Management Tools)，像是VirtualBox, VMware, Cloudera VM等，建構虛擬叢集(virtual cluster)，並安裝測試各項工具，藉以獲得實際操作經驗(hands-on experience)。

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	課程介紹和教學大綱(Course Introduction and Syllabus)	
2	大數據介紹(Introduction) . What is big data? Why big data? Examples of big data . The challenges and opportunities of big data	
3	General Purpose Platform I : Apache Hadoop and Ecosystem	
4	General Purpose Platform I : Apache MatReduce and algorithm design	
5	General Purpose Platform II: Apache Spark	
6	General Purpose Platform II: Apache Spark Ecosystem and algorithm design	
7	Big data storage architecture . Distributed nodes . Scale-out NAS . All-solid-satae drive (SSD) arrays . Object-based storage . DNA storage	
8	Big data storage systems . Distributed file systems and big data storage . Google GFS/Colossus and Apache HDFS . GlusterFS: Dependable Distributed File System	
9	No midterm. Independen study & presentation topics proposal instead.	
10	Big data systems for structured/semi-structured data . SQL, NoSQL, NewSQL, Distributed SQL . Apache HBase, Cassandra, CouchDB, Drill, Impala, Hive . Apache Sqoop	
11	Big data systems for structured/semi-structured data . Spark SQL, DataFrames, Datasets . MongoDB . Google BigQuery, Spanner, F1 . Presto	
12	Big graph processing systems . The challenges of big graph processing . Pregel family of systems(BSP, Pregel, Giraph)	
13	Big graph processing systems . GraphLab family of systems(GraphLab, PowerGraph, GraphChi) . Spark GraphX, GraphFrames . Neo4j graph database . Titan distributed graph database	

14	Big data stream processing systems . The challenges of big data streaming . Apache Storm, Samza, Flink . Apache SAMOA . Spark Streaming, Structured Streaming . Streaming data and data lake	
15	Big data pipelining, ETL(Extract, Transform, and Load), workflow and orchestration tools . Apache Airflow . Apache Kafka . Apache NiFi	
16	Big data pipelining, ETL(Extract, Transform, and Load), workflow and orchestration tools . Dagster . Prefect . Luigi . Google Cloud Dataflow/Composer, AWS Glue, Azure Data Factory**	
17	Independent study 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:

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: ____________________

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									
期中考成績 Midterm Exam	15%		✓	✓	✓				Independent Study & Presentation
期末考成績 Final Exam	25%				✓	✓			
作業成績 Homework and/or Assignments	35%		✓						
其他 Miscellaneous (學期計畫(Term Project))	25%		✓		✓	✓			

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

學期計畫需要對老師展演解說(Need to demo the term project to the instructor.)

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.)

無指定教科書，參考資料詳見課程講義.
(No textbook required. Check the class notes for references.)

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.)

所有教學講義資料都將公布在e-Learning課程網頁上。
所有作業繳交，也是透過e-Learning課程網頁。
老師的email信箱不穩定，繳交作業請務必在e-Learning網站上傳，勿透過email寄送。

Online class link: (if needed)
https://teams.microsoft.com/l/meetup-join/19%3aZLJ-YBvtT0PPw_YtdflYu5JV4C_K5bYsWcGP4d0tOus1%40thread.tacv2/1712625387041?context=%7b%22Tid%22%3a%22edba3211-8174-4411-b089-357c588fa127%22%2c%22Oid%22%3a%22e83708da-2e73-4b78-a037-e2bbca1f4d94%22%7d

其他補充說明（Supplemental instructions）