ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG MÔN HỌC

1. THÔNG TIN CHUNG (General information)

Tên môn học (tiếng Việt):	Công nghệ Big data
Tên môn học (tiếng Anh):	Big data technology
Mã môn học:	IE212
Thuộc khối kiến thức:	Đại cương ☐; Cơ sở nhóm ngành ☐;
	Cơ sở ngành ☐; Chuyên ngành ☐; Tốt nghiệp ☒
Khoa, Bộ môn phụ trách:	Khoa học và kỹ thuật thông tin
Giảng viên biên soạn:	**PGS.TS. Đỗ Phúc**
	Email: phucdo@uit.edu.vn
	Ths. Huỳnh Công Việt Ngữ
	Email: nguhcv@uit.edu.vn
Số tín chỉ:	4
Lý thuyết:	3
Thực hành:	1
Tự học:	
Môn học tiên quyết:	Không có
Môn học trước:	Cơ sở dữ liệu, Xác xuất thống kê

2. MÔ TẢ MÔN HỌC

Môn học giới thiệu tổng quan về khái niệm, đặc trưng cũng như những thách thức của Big data (Khả năng phân tích, dự đoán nhằm trích xuất một giá trị lớn hơn từ dữ liệu). Giới thiệu một số phương pháp và công cụ phổ biến để khai thác và quản lý Big data (Hadoop, MapReduce và Spark).

3. MỤC TIÊU MÔN HỌC (Course goals)

Ký hiệu	Mục tiêu môn học	Chuẩn đầu ra trong CTĐT
G1	Có khả năng phân tích, xử lý một vấn đề cụ thể liên quan đến Big data	LO3, LO5
G2	Biết sử dụng Hadoop - HDFS đê lưu trữ dữ liệu lớn	LO3, LO5, LO10
G3	Biết sử dụng mô hình MapReduce và Spark đê phân tích Big data	LO3, LO5, LO10
G4	Có khả năng sử dụng ngôn ngữ lập trình Java để phân tích dữ liệu lớn	LO3, LO5, LO10
G5	Có khả năng triển khai ứng dụng big data trong thực tế	LO10

4. CHUẨN ĐẦU RA MÔN HỌC (Course learning outcomes)

CĐRMH	Mô tả CĐRMH (Mục tiêu cụ thể)	Mức độ giảng dạy
G1.1 (LO3)	Trình bày được các khái niệm cũng như các đặc trưng cơ bản liên quan đến Big data.	I
G1.2 (LO3, LO5)	Có khả năng phân tích, đánh giá các vấn đề liên quan đến Big data.	IT
G2(LO3, LO5, LO10)	Có khả năng sử dụng HDFS để lưu trữ dữ liệu lớn trong môi trường Hadoop	ITU
G3.1 (LO3, LO5, LO10)	Có khả năng sử dụng Hadoop-MapReduce để phân tích và xử lý Big data.	ITU
G3.2 (LO3, LO5, LO10)	Có khả năng sử dụng Hadoop-Spark để phân tích và xử lý Big data.	ITU
G4 (LO3, LO5, LO10)	Có khả năng sử dụng ngôn ngữ lập trình Java để phân tích dữ liệu lớn	ITU
G5 (LO10)	Có khả năng triển khai ứng dụng big data trong thực tế	ITU

5. NỘI DUNG MÔN HỌC, KẾ HOẠCH GIẢNG DẠY (Course content, lesson plan)

a. Lý thuyết

Buổi học(3 tiết)	Nội dung	CĐRMH	Hoạt động dạy và học	Thành phần đánh giá
Buổi 1	**G**iới thiệu về khái niệm và một số kỹ thuật khai phá dữ liệu		Dạy: thuyết giảng Tự học: Đọc sách : Data Mining (Concepts and Techniques)	

Buổi 2	**Giới thiệu về Big data:** +) Khái niệm +) Các nét đặc trưng +) Nguồn hình thành +) Thử thách	G1	Dạy: Thuyết giảng	A2, A4
Buổi 3	**Hadoop:** +) Giới thiệu về mô hình GFS(Google File System) +) Lịch sử Hadoop +) Giải pháp Hadoop cho việc quản lý và khai thác Big data	G1	Dạy: Thuyết giảng	A2, A4
Buổi 4	**Hadoop:** +) Hệ thống file lưu trữ và quản lý của Hadoop: HDFS (Hadoop Distributed FileSystem)	G2	Dạy: Thuyết giảng	A2, A4
Buổi 5	**Hadoop:** +) Yarn +) Hadoop I/O	G2	Dạy: Thuyết giảng	A2, A4
Buổi 6	Giới thiệu khả năng dùng Java để viết chương trình xử lý big data (làm việc với MapReduce và Spark).	G4	Dạy: Thuyết giảng	A2, A4
Buổi 7, 8	**MapReduce (MR):** +) Giới thiệu về mô hình MR +) Cách thức phát triển một ứng dụng MR +) Xây dựng ứng dụng tiêu biểu phân tích Big data trên các tập dữ liệu mẫu có sẵn. Ví dụ: gen sinh học DNA, dữ liệu văn bản (text) hoặc dự liệu chứng khoán	G3.1, G4, G5	Dạy: Thuyết giảng	A2, A4
Buổi 9	**Spark:** +) Giới thiệu về Apache Spark +) Các tiềm năng và thử thách của Spark trong lĩnh vực "Khoa học dữ liệu (Data Science)"	G3.2, G4, G5	Dạy: Thuyết giảng	A2, A4
Buổi 10	**Spark:** +) Tìm hiểu về RDDs	G3.2, G4, G5	Dạy: Thuyết giảng	A2, A4
Buổi 11, 12, 13	**Spark:** +) Cách thức phát triển một ứng dụng toàn diện trong việc lưu trữ và phân tích dữ liệu +) Xây dựng ứng dụng tiêu biểu phân tích Big data trên các tập dữ liệu mẫu có sẵn(ví dụ: gen sinh học DNA, dữ liệu văn bản (text) hoặc dữ liệu chứng khoán)	G3.2, G4, G5	Dạy: Thuyết giảng	A2, A4

Buổi 14	**Ứng dụng Big data**	G3.2, G4, G5	Dạy: Thuyết giảng	A2, A4
Buổi 15	**Ôn tập tổng kết**	G1, G2, G3, G4	Dạy: Thuyết giảng	A2, A4

b. Thực hành:

Buổi học(3 tiết)	**Nội dung**	**CĐRMH**	**Hoạt động dạy và học**	**Thành phần đánh giá**
Buổi 1	**Hadoop:** **Cách cài đặt và hoạt động:** +) Standalone +) Pseudo-Distributed +) Fully-Distributed	G2	Dạy: Thực hành	A4
Buổi 2	**Hadoop:** Cách vận hành	G2	Dạy: Thực hành	A4
Buổi 3, 4, 5	**Hadoop- MapReduce** Lưu trữ và phân tích dữ liệu cơ bản với mô hình Hadoop-MapReduce Lập trình Java với Hadoop MapReduce	G2, G3.1, G4, G5	Dạy: Thực hành	A4
Buổi 6	**Spark:** Cài đặt và cách vận hành	G2, G3.2, G4, G5	Dạy: Thực hành	A4
Buổi 7,8,9, 10	**Spark:** Lưu trữ và phân tích dữ liệu với mô hình Hadoop-Spark Lập trình Java với Hadoop Spark	G2, G3.2, G4, G5	Dạy: Thực hành	A4

6. ĐÁNH GIÁ MÔN HỌC (Course assessment)

Thành phần đánh giá	**CĐRMH**	**Tỷ lệ (%)**
A1. Chuyên cần		**10%**
A2.Quá trình(kiểm tra, thuyết trình, báo cáo, …)	G1, G2, G3, G4, G5	**30%**
A3. Thi giữa kỳ		**0%**
A4. Đồ án cuối kỳ	G1, G2, G3, G4, G5	**60%**

a. Rubric của thành phần đánh giá A1

	Giỏi (4đ)	Khá (3đ)	TB (2đ)	Yếu (1đ)	Kém (0đ)
Tham gia đầy đủ các buổi học	- Tổng số buổi học: 14-15	- Tổng số buổi học: 12-13	- Tổng số buổi học: 10-11	- Tổng số buổi học: 8-9	- Tổng số buổi học: <=7

b. Rubric của thành phần đánh giá A2

	Giỏi (4đ)	Khá (3đ)	TB (2đ)	Yếu (1đ)	Kém (0đ)
Thực hiện đầy đủ và chính xác các bài kiểm tra	- Tổng điểm từ 80%	- Tổng điểm đạt từ 70 - 79%	- Tổng điểm đạt từ 60 - 69%	- Tổng điểm đạt từ 50 - 59%	- Tổng điểm dưới 50%

c. Rubric của thành phần đánh giá A4

	Giỏi (4đ)	Khá (3đ)	TB (2đ)	Yếu (1đ)	Kém (0đ)
- Xây dựng ứng dụng để phân tích dữ liệu lớn với mô hình Hadoop-MapReduce	- Ứng dụng thể hiện được tất cả các phần của kết quả phân tích dữ liệu	- Kết quả phân tích 60-79%	- Kết quả phân tích 40-59%	Kết quả đạt được <40%	Không thực hiện
- Xây dựng ứng dụng để phân tích dữ liệu lớn với mô hình Hadoop-Spark	- Ứng dụng thể hiện được tất cả các phần của kết quả phân tích dữ liệu	- Kết quả phân tích 60-79%	- Kết quả phân tích 40-59%	Kết quả đạt được <40%	Không thực hiện
- Viết báo cáo kết quả đạt được và so sánh hiệu suất giữa 2 phương pháp thực hiện	Báo cáo thể hiện được tất cả các khía cạnh phân tích và lập bảng so sánh chi tiết đánh giá theo yêu cầu	Kết quả đạt được 60-79%	Kết quả đạt được 40-59%	Kết quả đạt được <40%	Không thực hiện

7. QUY ĐỊNH CỦA MÔN HỌC (Course requirements and expectations)

- Tuân thủ các quy định của môn học
- Cách thức hoạt động và làm việc nhóm trên lớp.
- Phương pháp học tập của các bạn sinh viên tại lớp, về nhà.

8. TÀI LIỆU HỌC TẬP, THAM KHẢO

Giáo trình

1. Tom White(2015). Hadoop The Definitive Guide. Published by O' Reilly Media, Inc., Gravenstein Highway North, Sebastopol, CA 95472.
2. Holden Karau, Andy Kowinski and Matei Zaharia(2014). Learning Spark. Published by O' Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
3. Jiawei Han, Micheline Kamber, Jian Pei (2012). Data mining Concepts and Techniques. Published by Elservier, Inc., Waltham, MA 02451, USA.

Tài liệu tham khảo

1. Jeffrey Dean and Sanjay Ghemawat (2004). MapReduce:Simplified Data Processing on Large Clusters. OSDI 2004.
2. Rahul Beakta (2015). Big Data And Hadoop: A review Paper. BUEST, Baddi , RIEECE-2015
3. Matei Zaharia, Mosharaf Chowdhury, Micheal J.Franklin, Scott Shenker and Stoica . Spark: Cluster Computing with Working Sets. University of California, Berkeley.

9. PHẦN MỀM HAY CÔNG CỤ HỔ TRỢ THỰC HÀNH

1. Apache (June 08, 2017). Apache Hadoop 2.8.1
2. Apache (July 11, 2017). Apache Spark 2.2.0

Tp.HCM, ngày …… tháng …… năm ……

Trưởng khoa/bộ môn	**Giảng viên biên soạn**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)
	PGS.TS Đỗ Phúc