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 phân tích dữ liệu lớn
Tên môn học (tiếng Anh):	Big data analysis technologies
Mã môn học:	IE405
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 suất thống kê

2. MÔ TẢ MÔN HỌC

Môn học giới thiệu các kiến thức và các công nghệ phân tích dữ liệu lớn nhằm tìm kiếm tri thức từ dữ liệu lớn hỗ trợ tiến trình ra quyết định.

Môn học cung cấp khái niệm về dữ liệu lớn bao gồm 5 đặc điểm được viết tắt là 5V và các công cụ, kỹ thuật để lưu trữ và phân tích dữ liệu lớn như HDFS, MapReduce, Apache Spark,Mahout, hệ cơ sở dữ liệu NoSQL. Môn học còn giới thiệu cách dùng ngôn ngữ Python, Java, Scala để phân tích dữ liệu lớn. Cuối cùng, môn học giới thiệu vài ứng dụng của big data trong thực tiễn.

3. MỤC TIÊU MÔN HỌC (Course goals)

Ký hiệu	Mục tiêu môn học	Chuẩn đầu ra trong CTĐT
G1	Nắm vững kiến thức tổng quan, các nét đặc trưng và có cái nhìn sâu sắc về kỷ thuật phân tích Big data	LO3, LO5, LO10
G2	Có khả năng sử dụng HDFS và hệ cơ sở dữ liệu NoSQL để lưu trữ dữ liệu lớn	LO3, LO5, LO10
G3	Có khả năng sử dụng MapReduce, Apache Spark, Mahout để phân tích dữ liệu lớn	LO3, LO5, LO10
G4	Có khả năng sử dụng ngôn ngữ lập trình Python, Scala để phân tích dữ liệu lớn	LO3, LO5, 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 (LO10)	Có cái nhìn sâu sắc về các đặc điểm của các vấn đề kinh doanh đòi hỏi các kỹ thuật phân tích Big data	IT
G2.1 (LO3, LO5, LO10)	Có khả năng sử dụng HDFS để lưu trữ dữ liệu lớn trong môi trường Hadoop	ITU
G2.2 (LO3, LO5, LO10)	Có khả năng sử dụng hệ cơ sở dữ liệu NoSQL để lưu trữ dữ liệu lớn trong môi trường Hadoop	ITU
G3.1 (LO3, LO5, LO10)	Có khả năng sử dụng MapReduce và Mahout để phân tích chuyên sâu dữ liệu lớn	ITU
G3.2 (LO3, LO5, LO10)	Có khả năng sử dụng Apache Spark để phân tích chuyên sâu dữ liệu lớn	ITU
G4 (LO3, LO5, LO10)	Có khả năng sử dụng ngôn ngữ lập trình Python, Scala để phân tích dữ liệu lớn	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	+) Tổng quan và các đặc điểm đặc trưng Big data +) Góc nhìn thị trường và doanh nghiệp về tính cấp thiết của việc phân tích dữ liệu lớn	G1.1	Dạy: Thuyết giảng	A2, A4

Buổi 2	+) Các đặc điểm của các vấn đề kinh doanh phù hợp với giải pháp Big data +) Vai trò và trách nhiệm của doanh nghiệp trong việc triển khai các giải pháp dữ liệu lớn	G1.2	Dạy: Thuyết giảng	A2, A4
Buổi 3	+)Nhu cầu trong việc giám sát và quản trị dữ liệu lớn của doanh nghiệp +)Một số yêu cầu trong việc thiết kế phần cứng cho việc lưu trữ và phân tích dữ liệu lớn	G1.2	Dạy: Thuyết giảng	A2, A4
Buổi 4	+)Giới thiệu một số công cụ cũng như các kỹ thuật để quản lý và phân tích dữ liệu lớn	G1.2	Dạy: Thuyết giảng	A2, A4
Buổi 5	+)NoSQL – Mô hình lưu trữ và quản lý dữ liệu được áp dụng để phát triển ứng dụng dữ liệu lớn	G2	Dạy: Thuyết giảng	A2, A4
Buổi 6,7	+)Giới thiệu ngôn ngữ lập trình python, Scala	G4	Dạy: Thuyết giảng	A2, A4
Buổi 8,9	+)Mô hình lập trình Hadoop-Mapreduce	G2, G3, G4	Dạy: Thuyết giảng	A2, A4
Buổi 10, 11	+)Mô hình lập trình Hadoop-Spark	G2, G3, G4	Dạy: Thuyết giảng	A2, A4
Buổi 12, 13	+)Mô hình phân tích đồ thị trong một số vấn đề trong kinh doanh	G2, G3, G4	Dạy: Thuyết giảng	
Buổi 14, 15	+)Ôn tập và giới thiệu một số ứng dụng phân tích Big data trong thực tiễn	G2, G3, G4	Dạy: Thuyết giảng	

b. Thực hành:

Buổi học(3 tiết)	**Nội dung**	**CĐRMH**	**Hoạt động dạy và học**	**Thành phần đánh giá**
Buổi 1	+)Cài đặt và vận hành Hadoop	G2	Dạy: Thực hành	A2, A4
Buổi 2-5	**Hadoop- MapReduce:** +) Lưu trữ và phân tích dữ liệu với mô hình Hadoop- NoSQL-MapReduce với Python	G2, G3.1, G4	Dạy: Thực hành	A2, A4
Buổi 6-10	**Hadoop- Spark:** +) Lưu trữ và phân tích dữ liệu với mô hình Hadoop - NoSQL-Spark với Python	G2, G3.2, G4	Dạy: Thực hành	A2, 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	**30%**
A3. Thi giữa kỳ		**0%**
A4. Đồ án cuối kỳ	G1, G2, G3, G4	**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 chuyên sâu 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 chuyên sâu 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	Kết quả đạt được 60-79%	Kết quả đạt được 40-59%	Kết quả đạt được <40%	Không thực hiện

	theo yêu cầu				

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. David Loshin (2013). Big data analytics. 225 Wyman Street, Waltham, MA 02451, USA
3. Holden Karau, Andy Kowinski and Matei Zaharia(2014). Learning Spark. Published by O' Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
4. Wes McKinney (2013). Python for data analysis. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

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.
4. Martin Odersky, Lex Spoon, Bill Venners (2010). Programming in Scala. P.O.Box 305, Walnut Creek, California 94597.

9. PHẦN MỀM HAY CÔNG CỤ HỖ TRỢ THỰC HÀNH

1. Apache (June 08, 2017). Apache Hadoop 2.8.1
2. Apache (July 11, 2017). Apache Spark 2.2.0
3. Python (August, 09, 2017) Python 3.4.7
4. Scala 2.12.3

Tp.HCM, ngày …… tháng …… năm ……

Trưởng khoa/bộ môn

(Ký và ghi rõ họ tên)

PGS.TS Đỗ Phúc

Giảng viên biên soạn

(Ký và ghi rõ họ tên)