BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC SAO ĐỎ
*****

ĐỀ CƯƠNG CHI TIẾT HỌC PHẦN
DỮ LIỆU LỚN - BIG DATA

Số tín chỉ: 03
Trình độ đào tạo: Đại học
Ngành đào tạo: Công nghệ thông tin

Năm 2020

TRƯỜNG ĐẠI HỌC SAO ĐỎ
KHOA: ĐIỆN TỬ-TIN HỌC

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập-Tự do-Hạnh phúc

ĐỀ CƯƠNG CHI TIẾT HỌC PHẦN

Trình độ đào tạo: Đại học
Ngành đào tạo: Công nghệ thông tin

1. Tên học phần: Dữ liệu lớn - Big data
2. Mã học phần: CNTT 209
3. Số tín chỉ: 3 (2, 1)
4. Trình độ cho sinh viên: Năm thứ tư
5. Phân bổ thời gian
- Lên lớp: 30 tiết lý thuyết, 30 tiết thực hành.
- Tự học: 90 giờ.

6. Điều kiện tiên quyết: Không.
7. Giảng viên

STT	Học hàm, học vị, họ tên	Số điện thoại	Email
1	ThS. Phạm Thị Hường	0972.306.806	PTHuong@saodo.edu.vn
2	ThS. Nguyễn Thị Ánh Tuyết	0972.384.332	NTATuyet@saodo.edu.vn

8. Mô tả nội dung của học phần

Học phần Dữ liệu lớn - Big data 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.

9. Mục tiêu và chuẩn đầu ra học phần

9.1. Mục tiêu

Mục tiêu học phần thỏa mãn mục tiêu của chương trình đào tạo:

Mục tiêu	Mô tả	Mức độ theo thang đo Bloom	Phân bổ mục tiêu học phần trong CTĐT
MT1	**Kiến thức**		
MT1.1	Trình bày phương pháp phân tích, xử lý một vấn đề cụ thể liên quan đến Big data.	2	[1.2.1.2b]
MT1.2	Trình bày cách sử dụng công cụ Hadoop-HDFS để lưu trữ, mô hình MapReduce và Spark để phân tích dữ liệu lớn.	2	[1.2.1.2b]
MT1.3	Minh họa cách triển khai ứng dụng Big data trong thực tế.	3	[1.2.1.2b]

Mục tiêu	Mô tả	Mức độ theo thang đo Bloom	Phân bổ mục tiêu học phần trong CTĐT
MT2	**Kỹ năng**		
MT2.2	Áp dụng công cụ Hbase, Hadoop-HDFS để lưu trữ, mô hình MapReduce và Spark để phân tích dữ liệu lớn.	3	[1.2.2.2]
MT2.3	Phân tích, tổng hợp, đánh giá các công cụ phân tích trong lĩnh vực xử lý dữ liệu lớn.	4	[1.2.2.2]
MT3	**Mức tự chủ và trách nhiệm**		
MT3.1	Nghiêm túc, tự giác, tích cực, khoa học, độc lập, cẩn thận và tuân thủ trong công việc.	3	[1.2.3.1]
MT3.2	Có năng lực giải quyết vấn đề trong lĩnh vực Big data.	4	[1.2.3.2]

9.2. Chuẩn đầu ra

Sự phù hợp của chuẩn đầu ra học phần với chuẩn đầu ra của chương trình đào tạo:

CĐR học phần	Mô tả	Thang đo Bloom	Phân bổ CĐR học phần trong CTĐT
CĐR1	**Kiến thức**		
CĐR1.1	Giải thích được khái niệm, các đặc trưng cơ bản liên quan đến Big data.	2	[2.1.4]
CĐR1.2	Phân tích được các bước lưu trữ dữ liệu lớn bằng công cụ Hbase, Hadoop-HDFS.	4	[2.1.4]
CĐR1.3	Phân tích được các bước phân tích dữ liệu lớn bằng mô hình MapReduce và Spark.	4	[2.1.4]
CĐR2	**Kỹ năng**		
CĐR2.1	Áp dụng công cụ, mô hình để lưu trữ, phân tích và triển khai được dữ liệu lớn.	3	[2.2.3]
CĐR2.2	Đánh giá, cải tiến phương pháp để đáp ứng các tình huống thực tế trong lĩnh vực xử lý dữ liệu lớn.	5	[2.2.4]
CĐR3	**Mức tự chủ và trách nhiệm**		
CĐR3.1	Nghiêm túc, tự giác, tích cực, khoa học, độc lập, cẩn thận, tuân thủ trong lập trình và thực tế công việc.	3	[2.3.1]
CĐR3.2	Định hướng, hướng dẫn và đưa ra kết luận liên quan đến công việc phân tích và xử lý dữ liệu lớn.	4	[2.3.2]

10. Ma trận liên kết nội dung với chuẩn đầu ra học phần

Chương	Nội dung học phần	Chuẩn đầu ra của học phần						
		CĐR1			CĐR2		CĐR3	
		CĐR 1.1	CĐR 1.2	CĐR 1.3	CĐR 2.1	CĐR 2.2	CĐR 3.1	CĐR 3.2
1	**Chương 1. Giới thiệu về Big data** 1.1. Khái niệm Big data 1.2. Các kiểu Big data 1.3. Các đặc trưng của Big data	x			x		x	
2	**Chương 2. Hbase cho hệ thống Big data** 2.1. Giới thiệu về Hbase 2.2. Các tính năng của Hbase 2.3. Mô hình của Hbase 2.4. Kiến trúc Hbase 2.5. Cách thức lưu trữ và tìm kiếm của Hbase		x		x		x	
3	**Chương 3. Apache Hadoop cho hệ thống Big data** 3.1. Giới thiệu về mô hình GFS 3.2. Lịch sử Hadoop 3.3. Giải pháp Hadoop cho việc quản lý và khai thác Big data 3.4. Hệ thống file lưu trữ và quản lý của Hadoop: HDFS (Hadoop Distributed FileSystem) 3.5. Yarn 3.6. Hadoop I/O		x		x		x	
4	**Chương 4. Mô hình lập trình Mapreduce**			x		x		x

Chương	Nội dung học phần	Chuẩn đầu ra của học phần						
		CĐR1			CĐR2		CĐR3	
		CĐR 1.1	CĐR 1.2	CĐR 1.3	CĐR 2.1	CĐR 2.2	CĐR 3.1	CĐR 3.2
	4.1. Giới thiệu về mô hình Mapreduce-MR 4.2. Các hàm chính của MapReduce 4.3. Hoạt động của MapReduce 4.4. Cách thức phát triển một ứng dụng MR 4.5. Xây dựng ứng dụng phân tích Big data trên các tập dữ liệu mẫu có sẵn							
5	**Chương 5. Apache Spark cho hệ thống Big Data** 5.1. Tổng quan về Apache Spark 5.2. Các thành phần của Apache Spark 5.3. Quản lý bộ nhớ của Apache Spark 5.4. Lập trình với RDD 5.5. Phát triển ứng dụng lưu trữ và phân tích dữ liệu lớn			x		x	x	x

11. Đánh giá học phần

11.1. Kiểm tra và đánh giá trình độ

Chuẩn đầu ra	Mức độ thành thạo được đánh giá bởi
CĐR1	Kiểm tra thường xuyên, bài tập thực hành, kiểm tra thực hiện nhiệm vụ về nhà, kiểm tra giữa học phần.
CĐR2	Bài tập thực hành, thực hiện nhiệm vụ về nhà, kiểm tra giữa học phần, thi kết thúc học phần.
CĐR3	Kiểm tra thường xuyên, kết quả thực hiện nhiệm vụ của cá nhân và theo nhóm, thi kết thúc học phần.

11.2. Cách tính điểm học phần: Tính theo thang điểm 10 sau đó chuyển thành thang điểm chữ và thang điểm 4.

STT	Điểm thành phần	Quy định	Trọng số	Ghi chú
1	Điểm kiểm tra thường xuyên; điểm đánh giá nhận thức và thái độ tham gia thảo luận; điểm đánh giá phần bài tập; điểm chuyên cần	01 điểm	20%	Điểm trung bình của các lần đánh giá
2	Điểm kiểm tra giữa học phần	01 điểm	30%	
3	Điểm thi kết thúc học phần	01 điểm	50%	

11.3. Phương pháp đánh giá

Học phần sử dụng phương pháp đánh giá điểm thành phần như sau:

- Kiểm tra thường xuyên; đánh giá nhận thức và thái độ tham gia thảo luận; đánh giá nhiệm vụ tự học; chuyên cần: Vấn đáp.

- Kiểm tra giữa học phần: Thực hành (01 bài kiểm tra, thời gian làm bài: 90 phút).
- Thi kết thúc học phần: Bảo vệ bài tập lớn (20 phút/chủ đề).

12. Yêu cầu học phần

- Tham gia tối thiểu 80% số tiết học trên lớp dưới sự hướng dẫn của giảng viên.
- Đọc và nghiên cứu tài liệu phục vụ học phần, hoàn thành các bài tập cá nhân và bài tập nhóm.
- Chủ động làm bài tập lớn theo hướng dẫn của giảng viên.
- Tham gia kiểm tra giữa học phần, thi kết thúc học phần.
- Dụng cụ học tập: Máy tính, vở ghi, bút,...

13. Tài liệu phục vụ học phần

- Tài liệu bắt buộc:

[1] - Trường Đại học Sao Đỏ (2020), *Giáo trình Dữ liệu lớn - Big data.*

-Tài liệu tham khảo:

[2] - By Krishna Rungta (2019), *Learn Hadoop in 1 Day*.

[3] - Apache HBase ™ Reference Guide, *Introduction to Basic Schema* Design by Amandeep Khurana, Version 1.4.11.

[4] - Tom White (2015), *Hadoop The Definitive Guide*. Published by O' Reilly Media, Inc., Gravenstein Highway North, Sebastopol, CA 95472.

[5] - Holden Karau Andy Konwinski Matei Zaharia Patrick Wendell (2015), *Learning Spark*. Published by O' Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

14. Nội dung chi tiết học phần và phương pháp dạy-học

TT	Nội dung giảng dạy	Số tiết	Phương pháp dạy-học	CĐR học phần
1	**Chương 1. Giới thiệu về Big data** **Mục tiêu chương:** Sau khi học xong chương này, sinh viên đạt được các yêu cầu cơ bản sau: - Giải thích được khái niệm Big data, các kiểu Big data và đặc trưng của Big data. - Minh họa được các kiểu dữ liệu, các đặc trưng của Big data **Nội dung cụ thể:** 1.1. Khái niệm Big data 1.1.1. Data 1.1.2. Big data 1.2. Các kiểu Big data 1.2.1. Có cấu trúc 1.2.2. Không có cấu trúc 1.2.3. Bán cấu trúc 1.3. Các đặc trưng của Big data 1.3.1. Dung lượng dữ liệu 1.3.2. Tốc độ dữ liệu 1.3.3. Đa dạng dữ liệu Bài thực hành số 1.	4 (2LT, 2TH)	**Thuyết trình; Tổ chức học theo nhóm; Thực hành trên máy tính** **- Giảng viên:** + Giải thích khái niệm, kiểu dữ liệu và đặc trưng của Big data. + Giao bài tập, nội dung thực hành cho cá nhân, các nhóm. + Hướng dẫn sinh viên thực hành, đánh giá, nhận xét. **- Sinh viên:** + Đọc trước tài liệu: [1]: Chương 1; [2]: Chương 1; [3]: Chương 2. + Lắng nghe, ghi chép, quan sát. + Làm bài tập cá nhân, theo nhóm trong [1]: Chương 1. + Thực hành bài thực hành số 1.	CĐR1.1; CĐR2.1; CĐR3.1.
2	**Chương 2. Hbase cho hệ thống Big data** **Mục tiêu chương:** Sau khi học xong chương này, sinh viên đạt được các yêu cầu cơ bản sau: - Phân tích được các tính năng của Apache Hbase, mô hình Hbase, kiến trúc và cách lưu trữ dữ liệu của Hbase. - Áp dụng được Hbase lưu trữ dữ liệu trong Big data. **Nội dung cụ thể:**	12 (6LT, 6TH)	**Thuyết trình; Tổ chức học theo nhóm; Thực hành trên máy tính** **- Giảng viên:** + Giải thích các tính năng và sử dụng của Hbase. + Nêu nội dung vấn đề cần giải quyết. + Giao bài tập, nội dung thực hành cho cá nhân và các nhóm. + Hướng dẫn sinh viên thực hành, đánh giá, nhận xét.	CĐR1.2; CĐR2.1; CĐR3.1.

TT	Nội dung giảng dạy	Số tiết	Phương pháp dạy-học	CĐR học phần
	2.1. Giới thiệu về Hbase 2.2. Các tính năng của Hbase 2.3. Mô hình của Hbase 2.4. Kiến trúc Hbase 2.5. Cách thức lưu trữ và tìm kiếm của Hbase 2.6. Ví dụ áp dụng Bài thực hành số 2-4.		**- Sinh viên:** + Đọc trước tài liệu: [1]: Chương 2; [3]: Các chương. + Lắng nghe, ghi chép, quan sát, thảo luận. + Làm bài tập theo nhóm trong [1]: Chương 2. + + Thực hành bài thực hành số 2-4.	
3	**Chương 3. Apache Hadoop cho hệ thống Big data** **Mục tiêu chương:** Sau khi học xong chương này, sinh viên đạt được các yêu cầu cơ bản sau: - Phân tích được các tính năng của Apache Hadoop, mô hình Hbase, quản lý và khai thác big data của Hbase. - Đánh giá, lựa chọn được giải pháp quản lý và khái thác dữ liệu trong big data. **Nội dung cụ thể:** 3.1. Giới thiệu về mô hình GFS (Google File System) 3.2. Lịch sử Hadoop 3.3. Giải pháp Hadoop cho việc quản lý và khai thác Big data 3.4. Hệ thống file lưu trữ và quản lý của Hadoop: HDFS (Hadoop Distributed FileSystem) 3.5. Yarn 3.6. Hadoop I/O Bài thực hành số 5 - 7.	12 (6LT, 6TH)	**Thuyết trình; Dạy học dựa trên vấn đề; Tổ chức cho sinh viên tranh luận; Tổ chức học theo nhóm; Thực hành trên máy tính** **- Giảng viên:** + Giải thích tính năng, cách sử dụng Hadoop. + Nêu vấn đề, hướng dẫn sinh viên giải quyết vấn đề. + Nêu nội dung tranh luận. + Giao bài tập, nội dung thực hành cho cá nhân, các nhóm. + Hướng dẫn sinh viên thực hành, đánh giá, nhận xét. **- Sinh viên:** + Đọc trước tài liệu: [1]: Chương 3; [4]: Các chương. + Lắng nghe, ghi chép, quan sát, tranh luận, phản biện và giải quyết các vấn đề. + Làm bài tập cá nhân, theo nhóm trong [1]: Chương 3. + + Thực hành bài thực hành số 5 - 7.	CĐR1.2; CĐR2.1; CĐR3.1.

TT	Nội dung giảng dạy	Số tiết	Phương pháp dạy-học	CĐR học phần
4	**Chương 4. Mô hình lập trình Mapreduce** **Mục tiêu chương:** Sau khi học xong chương này, sinh viên đạt được các yêu cầu cơ bản sau: - Phân tích được mô hình Mapreduce, các hàm chính của Mapreduce, hoạt động của Mapreduce, cách thức phát triển Mapreduce. - Đánh giá, lựa chọn được cách sử dụng Mapreduce trong xử lý Big data. **Nội dung cụ thể:** 4.1. Giới thiệu về mô hình Mapreduce-MR 4.2. Các hàm chính của MapReduce 4.3. Hoạt động của MapReduce 4.4. Cách thức phát triển một ứng dụng MR 4.5. Xây dựng ứng dụng phân tích Big data trên các tập dữ liệu mẫu có sẵn 4.6. Ví dụ áp dụng **Kiểm tra giữa học phần** Bài thực hành số 8 - 9.	12 (6LT, 4TH, 2KT)	**Thuyết trình; Dạy học dựa trên vấn đề; Tổ chức học theo nhóm; Thực hành trên máy tính** **- Giảng viên:** + Giải thích tính năng và cách sử dụng mô hình Mapreduce. + Nêu vấn đề, hướng dẫn sinh viên giải quyết vấn đề. + Giao bài tập, nội dung thực hành cho cá nhân, các nhóm. + Hướng dẫn sinh viên thực hành, đánh giá, nhận xét. **- Sinh viên:** + Đọc trước tài liệu: [1]: Chương 4; + Lắng nghe, ghi chép, quan sát và giải quyết các vấn đề. + Làm bài tập cá nhân, theo nhóm trong [1]: Chương 4. + Làm bài kiểm tra + Thực hành bài thực hành số 8 - 9.	CĐR1.3; CĐR2.2; CĐR3.2.
5	**Chương 5. Apache Spark cho hệ thống Big data** **Mục tiêu chương:** Sau khi học xong chương này, sinh viên đạt được các yêu cầu cơ bản sau: - Phân tích được các thành phần của Apache Spark, các thành phần của Apache	20 (10LT, 10TH)	**Thuyết trình; Dạy học dựa trên vấn đề; Tổ chức học theo nhóm; Thực hành trên máy tính** **- Giảng viên:** + Giải thích tính năng và cách sử dụng mô hình Spark. + Nêu vấn đề, hướng dẫn sinh viên giải quyết vấn đề.	CĐR1.3; CĐR2.2; CĐR3.1; CĐR3.2.

TT	Nội dung giảng dạy	Số tiết	Phương pháp dạy-học	CĐR học phần
	Spark, quản lý bộ nhớ và lập trình với RDD. - Đánh giá, lựa chọn được các công cụ vào phân tích xử lý dữ liệu lớn thực tế. **Nội dung cụ thể:** 5.1. Tổng quan về Apache Spark 5.2. Các thành phần của Apache Spark 5.3. Quản lý bộ nhớ của Apache Spark 5.4. Lập trình với RDD 5.4.1. Tổng quan 5.4.2. Tạo RDD 5.4.3. Hoạt động của RDD 5.5. Phát triển ứng dụng lưu trữ và phân tích dữ liệu lớn 5.6. Ứng dụng Big Data Bài thực hành số 10 - 14.		+ Giao bài tập, nội dung thực hành cho cá nhân, các nhóm. + Hướng dẫn sinh viên thực hành, đánh giá, nhận xét. **- Sinh viên:** + Đọc trước tài liệu: [1]: Chương 5; [5]: Các chương. + Lắng nghe, ghi chép, quan sát và giải quyết các vấn đề. + Làm bài tập cá nhân, theo nhóm trong [1]: Chương 5. + Thực hành bài thực hành số 10 - 14.	

Hải Dương, ngày 24 tháng 09 năm 2020

KT.HIỆU TRƯỞNG PHÓ HIỆU TRƯỞNG	KT.TRƯỞNG KHOA PHÓ TRƯỞNG KHOA	TRƯỞNG BỘ MÔN
TS. Nguyễn Thị Kim Nguyên	Phạm Văn Kiên	Phạm Văn Kiên