IT4931 Lưu trữ và xử lý dữ liệu lớn

Phiên bản: 2019.1.0

1. THÔNG TIN CHUNG

Tên học phần: Lưu trữ và xử lý dữ liệu lớn
(Big data storage and processing)

Mã số học phần: IT4931

Khối lượng: 3(3-1-0-6)
- Lý thuyết: 45 tiết
- BTL: 15 tiết
- Thí nghiệm: 0 tiết

Học phần tiên quyết: CSDL

Học phần học trước:

Học phần song hành:

2. MÔ TẢ HỌC PHẦN

Dữ liệu lớn yêu cầu cần có các cơ chế, kỹ thuật xử lý dữ liệu ở quy mô lớn và hiệu quả. Học phần này nhằm cung cấp cho người học các kiến thức về công nghệ lưu trữ dữ liệu lớn, các hệ quản trị cơ sở dữ liệu NoSQL, NewSQL, các nguyên lý xử lý dữ liệu song song, phân tán, theo khối, theo luồng, xử lý sự kiện phức tạp, quản lý luồng công việc. Bên cạnh đó, người học được làm quen và vận dụng các công nghệ xử lý dữ liệu lớn trên nền tảng Hadoop - Map Reduce, và Spark. Sau khi kết thúc học phần này người học có khả năng hiểu, lựa chọn, cài đặt, và vận hành các giải pháp lưu trữ và xử lý dữ liệu phù hợp dựa trên các kiến thức học được cho các bài toán ứng dụng cụ thể liên quan tới dữ liệu lớn.

3. MỤC TIÊU VÀ CHUẨN ĐẦU RA CỦA HỌC PHẦN

Sinh viên hoàn thành học phần này có khả năng:

Mục tiêu/CĐR	Mô tả mục tiêu/Chuẩn đầu ra của học phần	CĐR được phân bổ cho HP/ Mức độ (I/T/U)
[1]	[2]	[3]
M1	**Hiểu và có khả năng thiết kế và quản lý các hệ thống thông tin trong các tổ chức**	1.1.4; 2.3.3; 3.1.4
M1.1	Nhận diện và hiểu rõ các thành phần của hệ thống thông tin	[1.1.4] (I)
M1.2	Nhận diện, so sánh và phân loại được các dạng thông tin và hệ thống thông tin trong doanh nghiệp	[1.1.4] (T)
M1.3	Có khả năng thiết kế hệ thống thông tin hỗ trợ truyền tải và trình bày dữ liệu, thông tin và tri thức trong tổ chức	[2.3.3; 3.1.4] (TU)
M2	**Nhận diện và làm chủ được các cơ hội trên thị trường do công nghệ thông tin đem lại để phát triển tổ chức sẵn có và tạo ra các tổ chức mới**	1.1.4; 3.1.5; 4.1.4; 5.1.4
M2.1	Hiểu và vận dụng được các ứng dụng công nghệ thông tin đương đại nhằm hỗ trợ các hoạt động trong tổ chức	[1.1.4; 3.1.5] (T)
M2.2	Nhận diện được các tác động của công nghệ thông tin đối	[4.1.4; 5.1.4] (U)

	với tổ chức và môi trường hoạt động của tổ chức	
M3	**Nhận diện các xu hướng phát triển của công nghệ thông tin có khả năng hỗ trợ việc thay đổi các tổ chức**	1.4.5; 4.1.1; 4.1.5
M3.1	Chủ động tìm hiểu và nhận diện các ứng dụng công nghệ thông tin mới nhất	[4.1.1; 4.1.5] (T)
M3.2	Xác định được các cơ hội mà công nghệ thông tin đem lại để phát triển tổ chức sẵn có	[1.4.5] (U)

4. TÀI LIỆU HỌC TẬP

Giáo trình

Bộ slide bài giảng

Sách tham khảo

[1] Tiwari, Shashank. Professional NoSQL. John Wiley & Sons, 2011.

[2] Lam, Chuck. Hadoop in action. Manning Publications Co., 2010.

[3] Miner, Donald, and Adam Shook. MapReduce design patterns: building effective algorithms and analytics for Hadoop and other systems. " O'Reilly Media, Inc.", 2012.

[4] Karau, Holden. Fast Data Processing with Spark. Packt Publishing Ltd, 2013.

[5] Penchikala, Srini. Big data processing with apache spark. Lulu. com, 2018.

[6] White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.

[7] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.

[8] Cattell, Rick. "Scalable SQL and NoSQL data stores." Acm Sigmod Record 39.4 (2011): 12-27.

[9] Gessert, Felix, et al. "NoSQL database systems: a survey and decision guidance." Computer Science-Research and Development 32.3-4 (2017): 353-365.

[10] George, Lars. HBase: the definitive guide: random access to your planet-size data. " O'Reilly Media, Inc.", 2011.

[11] Sivasubramanian, Swaminathan. "Amazon dynamoDB: a seamlessly scalable non-relational database service." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.

[12] Chan, L. "Presto: Interacting with petabytes of data at Facebook." (2013).

[13] Garg, Nishant. Apache Kafka. Packt Publishing Ltd, 2013.

[14] Karau, Holden, et al. Learning spark: lightning-fast big data analysis. " O'Reilly Media, Inc.", 2015.

[15] Iqbal, Muhammad Hussain, and Tariq Rahim Soomro. "Big data analysis: Apache storm perspective." International journal of computer trends and technology 19.1 (2015): 9-14.

[16] Toshniwal, Ankit, et al. "Storm@ twitter." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.

[17] Lin, Jimmy. "The lambda and the kappa." IEEE Internet Computing 21.5 (2017): 60-66.

5. CÁCH ĐÁNH GIÁ HỌC PHẦN

Điểm thành phần	Phương pháp đánh giá cụ thể	Mô tả	CĐR được đánh giá	Tỷ trọng
[1]	[2]	[3]	[4]	[5]
A1. Điểm quá trình (*)	**Đánh giá quá trình**			**40%**
	A1.1. Bài tập lớn	Lập trình		**40%**
A2. Điểm cuối kỳ	**A2.1. Thi cuối kỳ**	Thi viết hoặc Trắc nghiệm		**60%**

** Điểm quá trình sẽ được điều chỉnh bằng cách cộng thêm điểm chuyên cần. Điểm chuyên cần có giá trị từ –2 đến +1, theo Quy chế Đào tạo đại học hệ chính quy của Trường ĐH Bách khoa Hà Nội.*

Khóa học tham khảo:

1. https://www.coursera.org/learn/nosql-database-systems
2. https://who.rocq.inria.fr/Vassilis.Christophides/Big/index.htm
3. https://www.coursera.org/learn/big-data-introduction?specialization=big-data
4. https://www.coursera.org/learn/big-data-integration-processing?specialization=big-data
5. https://www.coursera.org/learn/big-data-management?specialization=big-data
6. https://www.coursera.org/learn/hadoop
7. https://www.coursera.org/learn/scala-spark-big-data

6. KẾ HOẠCH GIẢNG DẠY

Tuần	Nội dung	CĐR học phần	Hoạt động dạy và học	Bài đánh giá
[1]	[2]	[3]	[4]	[5]
1	**Chương 1: Tổng quan về lưu trữ và xử lý dữ liệu lớn** 1. Giới thiệu về Dữ liệu lớn (các khái niệm, các ứng dụng tạo ra và sử dụng dữ liệu lớn, ...) 2. Bài toán lưu trữ dữ liệu lớn (tổ chức, lưu trữ và quản lý) 3. Bài toán xử lý dữ liệu lớn 4. Hiện trạng lưu trữ và xử lý dữ liệu lớn trong thực tiễn (các thách thức công nghệ)		Bài giảng và tài liệu [7]	
2	**Chương 2: Hệ sinh thái Hadoop (Hadoop ecosystem)** 1. Giới thiệu về Hadoop		Bài giảng và tài liệu [6], Chapter 1	

	2. Các thành phần của hệ sinh thái Hadoop (kiến trúc, cấp phát tài nguyên với YARN, MapReduce, quản lý công việc trong MapReduce, …) 3. Giới thiệu Hadoop trên các dịch vụ đám mây			
3	**Chương 3: Hệ thống tập tin phân tán Hadoop HDFS** 1. Giới thiệu về HDFS 2. Kiến trúc HDFS 3. Đọc, ghi tệp và tổ chức tệp trong HDFS 4. Kiểu dữ liệu khóa và kiểu dữ liệu giá trị 5. Nguyên lý vào/ra dữ liệu song song 6. Các định dạng lưu trữ dữ liệu phổ biến với HDFS		Bài giảng và tài liệu Chapter 2, Chapter 3	
4	**Chương 4: Cơ sở dữ liệu phi quan hệ NoSQL - phần 1** 1. Các cuộc cách mạng cơ sở dữ liệu 2. Tổng quan các mô hình dữ liệu phi quan hệ 3. Định lý CAP 4. Mô hình nhất quán dữ liệu sau cùng (Eventual consistency model) 5. Các mô hình dữ liệu và lưu trữ 6. Ngôn ngữ truy vấn dữ liệu và giao diện lập trình 7. Các cơ sở dữ liệu phi quan hệ phổ biến		Bài giảng và tài liệu [8], [9]	
5	**Chương 4: Cơ sở dữ liệu phi quan hệ NoSQL - phần 2** 1. Giới thiệu Amazon DynamoDB (hoặc Hbase, hoặc Cassandra, tùy chọn một hệ thống phổ biến) 2. Kiến trúc phân tán dữ liệu của Amazon DynamoDB (hoặc Hbase, hoặc Cassandra, tùy chọn). (thiết lập môi trường, shell, tạo bảng, quản lý bảng, …).		Bài giảng và tài liệu [10], [11]	
6	**Chương 4: Cơ sở dữ liệu phi quan hệ NoSQL - phần 3** 1. Xử lý truy vấn SQL cho dữ liệu lớn (Hive). 2. Công nghệ lưu trữ NewSQL và tính chất		Bài giảng và tài liệu [12]	
7	**Chương 5: Hệ thống truyền thông điệp phân tán**		Bài giảng và tài liệu [13]	

	1. Giới thiệu và triển khai Apache Kafka 2. Kiến trúc phân tán 3. Mô hình publisher/consumer 4. Mô hình publisher/subscriber			
8	**Chương 6: Các kĩ thuật xử lý dữ liệu lớn theo khối - phần 1** 1. MapReduce 2. Một vài bài toán cơ bản trên MapReduce (Count, Sort, Pagerank)		Bài giảng và tài liệu [6], Chapter 2, 6, 7, 8	
9	**Chương 6: Các kĩ thuật xử lý dữ liệu lớn theo khối - phần 2** 1. Apache Spark 2. Tổ chức dữ liệu trong tập dữ liệu phân tán bền vững (Resilient distributed dataset) 3. Kiến trúc xử lý DAG (Directed Acyclic Graph) 4. Lập trình trên Spark dataframe		Bài giảng và tài liệu [14]	
10	**Chương 7: Các kĩ thuật xử lý luồng dữ liệu lớn** 1. Xử lý luồng với Spark streaming 2. Apache storm		Bài giảng và tài liệu [15] [16]	
11	**Chương 8: Kiến trúc dữ liệu lớn** 1. Kiến trúc dữ liệu lớn Lambda 2. Kiến trúc dữ liệu lớn Kappa		Bài giảng và tài liệu [17]	
12	**Chương 9: Phân tích dữ liệu lớn** 1. Một vài giải thuật phân tích dữ liệu cơ bản trên dữ liệu lớn 2. Spark ML		Bài giảng và tài liệu [14], Chapter 11	
13	**Báo cáo Bài tập lớn**			
14	**Báo cáo Bài tập lớn**			
15	**Báo cáo Bài tập lớn**			

7. QUY ĐỊNH CỦA HỌC PHẦN

(Các quy định của học phần nếu có)

8. NGÀY PHÊ DUYỆT: ……..

Chủ tịch Hội đồng **Nhóm xây dựng đề cương**

9. QUÁ TRÌNH CẬP NHẬT

Lần cập nhật	Nội dung điều chỉnh	Ngày tháng được phê duyệt	Áp dụng từ kỳ/khóa	Ghi chú
1	……			
2	……			