IT4931?? Lưu trữ và xử lý dữ liệu lớn

Bigdata Storage and Processing

Version: 2019.10.03

1. THÔNG TIN CHUNG
GENERAL INFORMATION

Tên học phần:	Lưu trữ và xử lý dữ liệu lớn
Course name:	*Bigdata storage and processing*
Mã học phần Code:	IT4931
Khối lượng Credit:	3(3-1-0-6) - Lý thuyết - Lecture: 45 hours - Bài tập - Exercise: 15 hours (If capstone project is used, please indicate clearly) - Thí nghiệm - Experiments: 0 hours
Học phần tiên quyết Prerequisite:	IT3090 Cơ sở dữ liệu IT3090 Database
Học phần học trước Prior course:	No
Học phần song hành Paralell course:	No

2. MÔ TẢ HỌC PHẦN - COURSE DESCRIPTION

Bigdata requires having many mechanisms, and techniques to process data at very large scale with efficiency. This course aims to provide student knowledge related to bigdata storage, NoSQL, NewSQL databases, principles of parallel and distributed data processing, batch and streaming processing, and complex event processing. Besides, students also are introduced bigdata processing on Hadoop – MapReduce and Spark technologies. After this course, students have the ability of understanding, selecting, deploying, and manipulating storage, processing solutions based on achieved knowledge for practical bigdata problem.

3. MỤC TIÊU VÀ CHUẨN ĐẦU RA CỦA HỌC PHẦN

GOAL AND OUTPUT REQUIREMENT

Sinh viên hoàn thành học phần này có khả năng

After this course the student will obtain the followings:

Mục tiêu/CĐR Goal	**Mô tả mục tiêu/Chuẩn đầu ra của học phần Description of the goal or output requirement**	**CĐR được phân bổ cho HP/ Mức độ (I/T/U) Output division/ Level (I/T/U)**
[1]	**[2]**	**[3]**
M1	**Applying basic and advanced scientific knowledge to**	1.1.2, 1.1.4, 1.2.1,

	building bigdata storage and processing solutions	1.2.2, 1.2.3, 1.2.6
M1.1	Able to apply basic scientific knowledge to build big data storage and processing solutions	1.1.2 [U] 1.1.4 [U] 1.2.1 [U]
M1.2	Able to apply the core scientific knowledge including computer systems, algorithms and programming, databases, design analysis ... in developing technical solutions for bigdata storage and processing.	1.2.2 [U] 1.2.3 [T] 1.2.6 [T]
M2	**Understanding and mastering knowledge of big data storage technology, NoSQL, NewSQL database management systems, big and complex data processing principles**	1.3.2, 1.3.3, 1.3.4
M2.1	Mastering information processing methods, architecture of distributed systems and information management techniques in distributed environments, Understanding about technologies for developing information systems. Having the ability to apply to the development of information systems serving organizations and businesses in storage, search and processing of information.	1.3.2 [T] 1.3.3 [T]
M2.2	Understanding and proficient use of programming tools and languages, development frameworks and common application architecture in building business applications, operating on different computing platforms	1.3.4 [I]
M3	**Applying teamwork skills, organization, coordination, effective team management, communication skills in foreign languages to practice.**	3.1.1, 3.1.2, 3.1.3, 3.1.4, 3.3.4
M3.1	Actively participating as well as being able to form a suitable team for assigned tasks	3.1.1 [IU]
M3.2	Organizing group activities	3.1.2 [IU]
M3.3	Managing group operation process	3.1.3 [IU]
M3.4	Having ability to cooperate, coordinate with other team members to solve problems	3.1.4 [IU]
M3.5	English reading skills	3.3.4 [U]

4. TÀI LIỆU HỌC TẬP

Study material

[1] Lecture slides

Reference book

[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 - EVALUATION

Điểm thành phần **Module**	**Phương pháp đánh giá cụ thể** **Evaluation method**	**Mô tả** **Detail**	**CĐR được đánh giá** **Output**	**Tỷ trọng** **Percent**
[1]	**[2]**	**[3]**	**[4]**	**[5]**
A1. Điểm quá trình **Mid-term (*)**	**Đánh giá quá trình** **Progress**			**40%**
	A1.1. Bài tập nhóm Capstone Project	Programming, Demo and Presentation	M1.1-M1.2 M2.1-M2.2 M3.1-M3.5	**40%**
A2. Điểm cuối kỳ **Final term**	**A2.1. Thi cuối kỳ** **Final exam**	Thi viết Written exam	M2.1-M2.2 M3.1-M3.5	**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.*

The evaluation about the progress can be adjusted with some bonus. The bonus should belong to [-2, +1], according to the policy of Hanoi University of Science and Technology.

Khóa học tham khảo

Reference course:

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 - SCHEDULE

Tuần Week	Nội dung Content	CĐR học phần Output	Hoạt động dạy và học Teaching activities	Bài đánh giá Evaluated in
[1]	[2]	[3]	[4]	[5]
1	**Chapter 1: Overview of big data storage and processing** 1. Introduction to Big Data (concepts, applications that create and use big data, ...) 2. Big data storage problem (organization, storage and management) 3. Big data processing problem. 4. Current situation of big data storage and processing (technological challenges)	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [7]	A1 A2
2	**Chapter 2: Hadoop ecosystem** 1. Introduction to Hadoop 2. Components of Hadoop ecosystem (architecture, resources allocation with YARN, MapReduce, job management in MapReduce, ...) 3. Introducing Hadoop on cloud services	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [6], Chapter 1	A1 A2
3	**Chapter 3: Hadoop distributed file system (HDFS)** 1. Introduction to HDFS 2. HDFS architecture 3. Read, write files and organize files in HDFS 4. Key data type and value data type 5. The principle of parallel input / output 6. Popular data storage format with HDFS	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [6], Chapter 2, 3	A1 A2
4	**Chapter 4: NoSQL relational database - part 1** 1. The database revolution 2. Overview of non-relational data models 3. The CAP theorem 4. Eventual consistency model	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [8], [9]	A1 A2

	5. Data models and storage 7. Data query language 8. Popular non-relational databases			
5	**Chapter 4: NoSQL relational database - part 2** 1. Introducing Amazon DynamoDB (or Hbase, or Cassandra - optional). 2. Data distribution architecture of Amazon DynamoDB (or Hbase, or Cassandra, optional). (environment settings, shell, table creation, table management, ...).	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [10], [11]	A1 A2
6	**Chapter 4: NoSQL relational database - part 3** 1. Handling SQL queries for big data (Hive). 2. NewSQL storage technology and properties	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [12]	A1 A2
7	**Chapter 5: Distributed messaging system** 1. Introduction and deployment of Apache Kafka 2. Distributed architecture of distributed messaging system 3. Publisher/consumer model 4. Publisher/subscriber model	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [13]	A1 A2
8	**Chapter 6: Mass data processing techniques - part 1** 1. MapReduce 2. Several basic problems on MapReduce (Count, Sort, Pagerank)	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [6], Chapter 2, 6, 7, 8	A1 A2
9	**Chapter 6: Mass data processing techniques - part 2** 1. Apache Spark 2. Organizing data in resilient distributed dataset 3. DAG processing architecture (Directed Acyclic Graph) 4. Programming on Spark dataframe	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [14]	A1 A2
10	**Chapter 7: Large data flow processing techniques** 1. Process streams with Spark streaming 2. Apache storm	M1.1 M1.2 M2.1 M2.2 M3.5	Lecture and reference [15], [16]	A1 A2
11	Chapter 8: Big data architecture 1. Lambda big data architecture 2. Kappa big data architecture	M1.1 M1.2 M2.1 M2.2	Lecture and reference [17]	A1 A2
12	**Chapter 9: Big data analysis** 1. Several basic data analysis algorithms on big data	M1.1 M1.2 M2.1	Lecture and reference [14], Chapter 11	A1 A2

	2. Spark ML	M2.2		
13	Presentation capstone project	M3.1 M3.2 M3.3 M3.4		A1
14	Presentation capstone project	M3.1 M3.2 M3.3 M3.4		A1
15	Presentation capstone project	M3.1 M3.2 M3.3 M3.4		A1

7. QUY ĐỊNH CỦA HỌC PHẦN - COURSE REQUIREMENT

(The specific requirements if any)

8. NGÀY PHÊ DUYỆT - DATE: ……..

Chủ tịch hội đồng
Committee chair

Nhóm xây dựng đề cương
Course preparation group
Nguyen Binh Minh, Tran Viet Trung, Nguyen Ba Ngoc, Nguyen Kim Anh, Tran Hai Anh

9. QUÁ TRÌNH CẬP NHẬT - UPDATE INFORMATION

STT No	**Nội dung điều chỉnh Content of the update**	**Ngày tháng được phê duyệt Date accepted**	**Áp dụng từ kỳ/ khóa Applicable from**	**Ghi chú Note**
1	……			
2	……			