BIG DATA – MSIS405

Syllabus

1. GENERAL INFORMATION:

Instructor name: Email:

Credit: 3 (3 lecture).

Prerequisites:

2. COURSE INFORMATION:

- *Course description*:

Storage, retrieval, analysis, and knowledge discovery using Big Data has made significant inroads in several domains in industry, research, and academia and look at the dominant software systems and algorithms for coping with Big Data. Topics covered include large-scale non-traditional data storage frameworks including graph, key-value, and column-family storage systems; data stream analysis algorithms; large scale anomaly detection; information diffusion; and recommendation algorithms. The course will involve hands-on proramming assignments and a term-project using real-world datasets.

3. BOOK AND MATERIALS:

- Required textbook:

Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data by Paul C. Zikopoulos, Chris Eaton, McGraw Hill Publisher, 2011.

- Other materials:

1. Social Network Analysis (4th Ed.) by John Scott, SAGE Publications Ltd, 2017

2. Mining of massive datasets, Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, online version and research articles related.

4. GRADING PROCEDURES:

Assignments, Class attendance/participation: 50%

Final Project: 50%

5. COURSE OUTLINE:

Topics:
- Polyglot persistence
- Key-value storage systems
- Column-family storage systems
- Graph storage systems
- Algorithms for detecting similar items
- Recommandation systems
- Data stream analysis algorithms
- Link Analysis algorithms
- Clustering algorithms
- Detecting frequent items

Week	Topic
1	**Part 1. Introduction to Big Data** Week 1: Topics Introduction to Big Data Revisit useful technologies and concepts Readings : J. Ginsberg, et al., "Detecting influenza epidemics using search engine query data" Nature 457 pp. 1012 ~ 1014, February 2009 Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, "The Google file system" Proceedings of SOSP 2003: 29-43
2	Week 2: Topics Distributed File System: Continued Readings : Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, "The Google file system" Proceedings of SOSP 2003: 29-43
3	**Part 2. Data Storage Models** Week 3: Topics NoSQL stands for "Not Only SQL". Why NoSQL? Data Consistency Distributed Hashtable Key-Value storage systems (Amazon's Dynamo) Readings : Giuseppe DeCandia, et al., "Dynamo: Amazon's Highly Available Key-value Store," Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles, pp. 205-220
4	Week 4: Topics Key-Value storage systems-continued (Amazon's Dynamo) Readings : Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan,"Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications" Proc. 2001 SIGCOMM, Mar. 2001, pp.149-160
5	Week 5: Topics Column-Family storage models (Google's BigTable) Readings :

	Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber, "Bigtable: A Distributed Storage System for Structured Data" OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006
6	Week 6: Topics Document storage systems (Facebook's Cassandra) **Midterm 1**
7	Week 7: Topics Document storage systems (Facebook's Cassandra) Readings : Avinash Lakshman, Prashant Malik, "Cassandra: A Decentralized Structured Storage System" ACM SIGOPS Operation Systems Review, Vol. 44-(2), April 2010 pp. 35-40 Related links : Titan: http://thinkaurelius.github.io/titan/ Gremin: https://github.com/tinkerpop/gremlin/wiki Faunus: http://thinkaurelius.github.io/faunus/
8	Week 8: Topics Graph storage models Related links : Titan: http://thinkaurelius.github.io/titan/ Gremin: https://github.com/tinkerpop/gremlin/wiki Faunus: http://thinkaurelius.github.io/faunus/ Grzegorz Malewicz et. el. "Pregel: a system for large-scale graph processing" Proceeding SIGMOD '10 Proceedings of the 2010 ACM SIGMOD International Conference on Management of data Pages 135-146
9	**Part 3. Scalable algorithms and Big Data Analytics** Week 9: Topics Recommendation systems with case studies of Amazon's Item-to-Item recommendation and Netfix Prize Readings : Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012 --Chapter 9 Greg Linden, Brent Smith, and Jeremy York, "Amazon.com Recommendations, Item-to-Item Collaborative Filtering" IEEE Internet Computing, 2003 Yehuda Keren, "Matrix Factorization Techniques For Recommender System", IEEE Computer 2009
10	Week 10: Topics Link Analysis with case studies of the PageRank algorithm and the Spam farm analysis Readings : Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012 --Chapter 3 and

	Chapter 5
11	Week 11: Topics Mining Data Streams Readings : Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012 -- Chapter 4
12	Week 12: Topics Mining Data Streams : Continued Readings : Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012 -- Chapter 4 Hadoop mini tutorial.
13	Week 13: Topics Advertising on the Web Readings : Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012 -- Chapter 8
14	Week 14: Topics Advertising on the Web Readings : Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012 -- Chapter 8
15	Week 15: Project Presentations

6. COURSE REQUIREMENTS:

There will be several individual or team presentations. (Each presentation as well as how a group functioned as a team will be graded subjectively by the instructor)

Each individual assignment is to be done independently. Students are encouraged to join in the class discussion and present their thoughts and ideas on the all distributed system problems.

Students are expected to attend all class sessions. Excused absences will be granted only in cases of illness, death, or other extreme family emergency. There is a grade penalty for excessive absences. Any request for an excused absence must be made in person and in writing.

7. ACADEMIC INTEGRITY POLICIES:

- Student may not use Vietnamese languague in class, or will be reduced 2% final marks
- Be punctual to come and leave the class.
- Maximum cancellation time per semester is 6 hours per class.

Instructor's Signature