جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Course Syllabus: Advanced Topics in Data Management - CS 341

Division	Computer, Electrical and Mathematical Sciences & Engineering
Course Number	CS 341
Course Title	Advanced Topics in Data Management
Academic Semester	Fall
Academic Year	2019/2020
Semester Start Date	08/25/2019
Semester End Date	12/10/2019
Class Schedule (Days & Time)	09:00 AM - 10:30 AM | Mon Thu

Instructor(s)

Name	Email	Phone	Office Location	Office Hours
Panagiotis Kalnis	panos.kalnis@kaust.edu.sa	+966128080343	0111, 1, Al-Khawarizmi (bldg. 1)	I am available almost every day. Please email me for an appointment: panos.kalnis@kaust.edu.sa

Teaching Assistant(s)

Name	Email
n.a.	n.a.

Course Information

Comprehensive Course Description	The course will focus on Data Management and Analysis using Parallel and Distributed Systems. Topics will include: Distributed Hash Tables, Map-Reduce, Spark, Pregel, Graph processing, Large-scale Machine Learning, and Cloud Databases. Every lecture will focus on a research paper. All students will have to read the paper and write a short review (roughly 150 words of summary plus pros and cons of the paper). One student will present the paper (roughly 45min) and then the whole class will discuss the work. Each student must complete a substantial programming project.
Course Description from Program Guide	Topics in Data Management will be analyzed and discussed. Students will engage in research and project presentations. Topics will vary by semester.
Goals and Objectives	The students will learn the state-of-the art in using very large distributed and parallel architectures to process Big Data and perform Machine Learning tasks.
Required Knowledge	Students must have taken CS245 or CS220, or equivalent course and must have excellent programming experience in C/C++ and Linux.
Reference Texts	The course will be based on research papers. All papers can be found on the course's wiki: http://cs341.pbworks.com. No textbook is needed. The library has several books on databases that can be used for reference.
Method of evaluation	**20.00**% - Homework /Assignments **20.00**% - Oral presentation **40.00**% - Course Project(s) **20.00**% - Active participation

Nature of the assignments	All students will have to read each paper and write a short review (roughly 150 words of summary plus achpros and cons of the paper). One student will present the paper (roughly 45min) and then the whole class will discuss the work. Each student must complete a substantial programming project.
Course Policies	Students may miss up to 2 lectures.
Additional Information	Refer to: http://cs341.pbworks.com

Tentative Course Schedule

(Time, topic/emphasis & resources)

Week	Lectures	Topic
1	Mon 08/26/2019 Thu 08/29/2019	Introduction Selection of project topic
2	Mon 09/02/2019 Thu 09/05/2019	[1] Distributed Query Processing in a Relational Data Base System, R.S. Epstein, M. Stonebraker, E. Wong, In Proc. of SIGMOD, pp. 169-180, 1978. [2] The Gamma Database Machine Project, D.J. DeWitt, S. Ghandeharizadeh, D.A. Schneider, A. Bricker, H. Hsiao, R. Rasmussen, IEEE TKDE, 2(1), pp. 44-62, 1990.
3	Mon 09/09/2019 Thu 09/12/2019	[3] Mariposa: a wide-area distributed database system, M. Stonebraker, P.M. Aoki, W. Litwin, A. Pfeffer, A. Sah, J. Sidell, C. Staelin, A. Yu, The VLDB Journal, 5(1), pp. 48-63, 1996. [4] Eddies: Continuously Adaptive Query Processing, R. Avnur, J.M. Hellerstein, In Proc. of SIGMOD, pp. 261-272, 2000.
4	Mon 09/16/2019 Thu 09/19/2019	[5] Chord: a scalable peer-to-peer lookup protocol for internet applications, I. Stoica, R. Morris, D. Liben-Nowell, D.R. Karger, M.F. Kaashoek, F. Dabek, H. Balakrishnan, IEEE/ACM Transactions on Networks, 11(1), pp. 17-32, 2003. [6] Database Cracking, S. Idreos, M. Kersten, S. Manegold, In Proc. of CIDR, 2007.
5	Mon 09/23/2019 Thu 09/26/2019	Saudi National Day
6	Mon 09/30/2019 Thu 10/03/2019	[9] The Google File System. S. Ghemawat, H. Gobioff, S.T. Leung, In Proc. of ACM Symposium on Operating Systems Principles (SOSP),pp. 29-43, 2003. [10] MapReduce: Simplified Data Processing on Large Clusters. J Dean and S. Ghemawat, Proc. of Symposium on Operating System Design and Implementation (OSDI), 2004.
7	Mon 10/07/2019 Thu 10/10/2019	[11] Bigtable: A distributed storage system for structured data, F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, In Proc. of USENIX-OSDI, pp. 205-218, 2006. [12] MapReduce and parallel DBMSs: friends or foes?. M. Stonebraker, D. Abadi, DJ. DeWitt, S. Madden, E. Paulson, A. Pavlo, and A. Rasin. Communications of the ACM. 53(1), pp. 64-71, 2010.
8	Mon 10/14/2019 Thu 10/17/2019	[13] Pig latin: a not-so-foreign language for data processing, C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins, In Proc. of SIGMOD, pp. 1099-1110, 2008. [14] Spark: Cluster Computing with Working Sets. Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica. In Proc. of HotCloud, 2010.
9	Mon 10/21/2019 Thu 10/24/2019	[15] Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In- Memory Cluster Computing. Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica. In Proc. of NSDI, 2012. [16] Shark: SQL and Rich Analytics at Scale. Reynold S. Xin, Joshua Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, Ion Stoica. In Proc. of SIGMOD, 2013.
10	Mon 10/28/2019 Thu 10/31/2019	[17] Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis, Tal Ben-Nun, Torsten Hoefler, arXiv.org, 2018 [18] A domain-specific architecture for deep neural networks. N.P. Jouppi, C. Young, N., D. Patterson. In Commun. ACM 61(9), pp. 50-59, 2018. (August 2018), 50-59.
11	Mon 11/04/2019 Thu 11/07/2019	[19] TensorFlow: A system for large-scale machine learning. M. Abadi et al., In Proc. of 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2016. [20] Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge. Y. Kang, et al. In Proc. of the Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS '17). pp. 615-629, 2017.
12	Mon 11/11/2019 Thu 11/14/2019	[21] Project Adam: building an efficient and scalable deep learning training system. Trishul Chilimbi, et al., In Proc of OSDI, pp. 571-582, 2014. [22] NoScope: optimizing neural network queries over video at scale. D. Kang, et al., Proc. VLDB Endow. 10(11), pp. 1586-1597, 2017.
13	Mon 11/18/2019 Thu 11/21/2019	[23] On Optimizing Operator Fusion Plans for Large-Scale Machine Learning in SystemML M. Boehm, et.al, PVDLB, 2018. [24] TVM: An Automated End-to-End Optimizing Compiler for Deep Learning, T. Chen et al., CoRR, 2018.
14	Mon 11/25/2019 Thu 11/28/2019	[25] Compressed linear algebra for large-scale machine learning, Ahmed Elgohary, Matthias Boehm, Peter J. Haas, Frederick R. Reiss, Berthold Reinwald, VLDBJ, 2018. [26] Low-Memory Neural Network Training: A Technical Report, Nimit Sharad Sohoni, Christopher Richard Aberger, Megan Leszczynski, Jian Zhang, Christopher Ré, arXiv.org, 2019
15	Mon 12/02/2019 Thu 12/05/2019	Project presentations
16	Mon 12/09/2019	Exams

Note

The instructor reserves the right to make changes to this syllabus as necessary.