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	2018/2019
Semester Start Date	08/26/2018
Semester End Date	12/11/2018
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 on Parallel and Distributed systems. Topics will include: Distributed Semi-joins, Distributed Hash Tables, CAN, Chord, the Map-Reduce framework, Spark, Pregel, Eddies, Graph processing, implementation of relational operators on multicore architectures, databases in the Cloud, and others. Every lecture will focus on one research papers. 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.
Required Knowledge	Students must have taken CS245 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**% - Active participation **40.00**% - Course Project(s) **20.00**% - Oral presentation **20.00**% - Homework /Assignments

Nature of the assignments	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 Policies	Students may miss up to 2 lectures.
Additional Information	

Tentative Course Schedule

(Time, topic/emphasis & resources)

Week	Lectures	Topic
1	Mon 08/27/2018 Thu 08/30/2018	Introduction Selection of project topic
2	Mon 09/03/2018 Thu 09/06/2018	[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 09/10/2018 Thu 09/13/2018	[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 09/17/2018 Thu 09/20/2018	[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 09/24/2018 Thu 09/27/2018	[7] Schism: a workload-driven approach to database replication and partitioning, C. Curino, E. Jones, Y. Zhang, and S. Madden. 2010. Proc. VLDB Endow. 3, 1-2, pp. 48-57, 2010. [8] C-Store: A Column-oriented DBMS, M. Stonebraker, D.J. Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E. O'Neil, P O'Neil, A. Rasin, N. Tran, S. Zdonik, In Proc. of VLDB, pp. 553-564, 2005.
6	Mon 10/01/2018 Thu 10/04/2018	[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 10/08/2018 Thu 10/11/2018	[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 10/15/2018 Thu 10/18/2018	[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 10/22/2018 Thu 10/25/2018	[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 10/29/2018 Thu 11/01/2018	[17] Spark SQL: Relational Data Processing in Spark. Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, Matei Zaharia. In proc of SIGMOD, 2015. [18] Pregel: a system for large-scale graph processing, G. Malewicz, M.H. Austern, A.J.C Bik, J.C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. In Proc. of SIGMOD, 2010.
11	Mon 11/05/2018 Thu 11/08/2018	SPRING BREAK
12	Mon 11/12/2018 Thu 11/15/2018	[19] GraphChi: large-scale graph computation on just a PC. Aapo Kyrola, Guy Blelloch, and Carlos Guestrin, In Proc of USENIX OSDI, 2012 [20] GraphX: Unifying Data-Parallel and Graph-Parallel Analytics. Reynold S. Xin, Daniel Crankshaw, Ankur Dave, Joseph E. Gonzalez, Michael J. Franklin, Ion Stoica. OSDI 2014. October 2014.
13	Mon 11/19/2018 Thu 11/22/2018	[21] Distributed GraphLab: a framework for machine learning and data mining in the cloud. Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M. Hellerstein. Proc. of the VLDB Endow. 5(8), 2012. [22] Mizan
14	Mon 11/26/2018 Thu 11/29/2018	[23] AdPart [24] ScaleMine
15	Mon 12/03/2018 Thu 12/06/2018	[25] Hideaki Kimura, Alkis Simitsis, Kevin Wilkinson: Janus: Transaction Processing of Navigation and Analytic Graph Queries on Many-core Servers [26] The TileDB Array Data Storage Manager, Stavros Papadopoulos, Kushal Datta, Samuel Madden, Timothy Mattson http://www.vldb.org/pvldb/vol10/p349-papadopoulos.pdf

16	Mon 12/10/2018	[27] An Experimental Comparison of Partitioning Strategies in Distributed Graph Processing, Shiv Verma, Luke Leslie, Yosub Shin, Indranil Gupta, http://www.vldb.org/pvldb/vol10/p493-verma.pdf [28] Distributed Join Algorithms on Thousands of Cores, Claude Barthels, Gustavo Alonso, Torsten Hoefler, Timo Schneider, Ingo Muller, http://www.vldb.org/pvldb/vol10/p517-barthels.pdf

Note

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