Advanced Database Systems

RUTGERS
Rutgers Business School
Newark and New Brunswick

26-198-641: Advanced Database Systems Fall 2022 Section 1: 1-WP-220 [Newark Campus] Wednesday, 10:00-12:50	**Dr. Joann J Ordille** Associate Professor of Practice Office: Levin 231 [Livingston Campus] Office: TBD [Newark Campus] Office Phone: 848-445-3243 (shared (Do not leave message on phone. I do not yet have the code for retrieving them.) jo531@scarletmail.rutgers.edu Office Hours: T,Th: 2:30 – 3:30 pm [Livingston Campus] T: 5: 20 – 5:50 pm [Livingston Campus] W: 1: 45 – 2:45 pm [Newark Campus] Office hours are in-person on the designated campus and virtual via Zoom. You can also make an appointment.

COURSE DESCRIPTION

This course focuses on research and applications in advanced database systems for Cloud and Big Data Computing. It provides an opportunity to learn about Cloud Computing and Advanced Database Systems and apply that learning on a popular cloud platform. The course topics include how database systems have addressed the four V's of Big Data: volume, variety, velocity and veracity. We also consider maintaining the virtue of our data, a fifth V if you will, by addressing issues of security, privacy, and social responsibility.

Advanced database research has produced a collection of powerful and successful NoSQL (Not Only SQL) database systems, each of which addresses the four V's. The course includes Amazon's DynamoDB and Google's Megastore as examples of key-value stores. Key-value stores form the foundation for fast, incrementally scalable, distributed processing of Internet shopping carts, user information, and product information. The course discusses Google's BigTable and Facebook's Cassandra as examples of wide-column databases. These databases support fast information storage and retrieval for search engines, personalization of services, analytics, and email. The course includes MongoDB as an example of a document database. MongoDB undergirds the high performance of many web sites and web applications. It is currently the most popular NoSQL database. Neo4j and Pregel are included as examples of graph databases that support analyzing social media relationships, transportation systems, disease outbreaks, and other graphs. Spark Streaming is our example of a popular system for processing data generated at high velocity such as data generated by sensors in the Internet of Things (IOT). We examine how these databases conform to the CAP Theorem by making tradeoffs between data consistency, availability, and resilience to network partitioning in order to achieve scale. We also explore how underlying technologies like MapReduce make these systems possible.

During Fall 2022, free access to Amazon Web Services (AWS), the Amazon Cloud Platform, is provided to students in this course as part of the AWS Academy Program.

COURSE MATERIALS

- **IMPORTANT:** The original resource for our readings, which provided free access to Association of Computing Machinery (ACM) members, has been discontinued. I've revised the reading list of required books and provide pointers for purchasing at a lower price. The books will also be available in the library. You do NOT need to join the ACM to obtain materials for this course.
- Required books:
    - Carpenter, J. & Hewitt, E. (2022). *Cassandra: the definitive guide* (2nd ed.). O'Reilly Media, Inc. The second edition is available used or in overstock at a much lower price from the third edition. The second edition is sufficient for our needs.
    - Damji, J., Lee, D., Wenig, B., & Das, T. (2020). *Learning Spark: lightning-fast big data analysis* (2nd ed.) O'Reilly Media, Inc. Available for rent on Amazon, as well as used and new from a variety of vendors.
    - Harrison, G. (2016). *Next generation databases: NoSQL, newSQL, and big data*. Apres. Look for it used or in overstock on the Internet for a much lower price. An electronic version can be rented from Amazon.
    - Perkins, L., Redmond, E., & Wilson, J. (2018). *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf. Consider buying it in electronic format direct from the publisher for a lower price.
- Recommended book:
    - Lin, J., & Dyer, C. (2010). Data-intensive text processing with MapReduce. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-177. Free access available at: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
- Articles in conferences proceedings, journals and professional publications are used in this course as described in the timetable below.
- Check Canvas (https://canvas.rutgers.edu/) and your Scarlet Mail Rutgers email account regularly for additional course materials.

PREREQUISITES

Students taking this course should have knowledge of relational database systems and experience in computer programming.

ACADEMIC INTEGRITY

I do NOT *tolerate cheating*. Students are responsible for understanding the RU Academic Integrity Policy (http://academicintegrity.rutgers.edu/). I will strongly enforce this Policy and pursue *all* violations. On all examinations and assignments, students must sign the RU Honor Pledge, which states, "On my honor, I have neither received nor given any unauthorized assistance on this examination or assignment." Failure to sign the honor statement will result in a zero for the examination or assignment. Don't let cheating or plagiarism destroy your hard-earned opportunity to learn. See business.rutgers.edu/ai for more details.

CLASSROOM CONDUCT

Research has shown that students learn better in a community with their peers. We hope to help you form that community by creating teams. These teams will participate in class in group activities. They will collaborate in reading and discussing research papers in preparation for class meetings. Teams will submit summaries of their discussions, or be required to ask or answer questions in class. Each team will also have the responsibility for presenting a set of papers for one of the classes. Teams will consult with me in advance of their presentation, and every member must take an active role in doing the presentation.

In class, we will sometimes have active review sessions. A series of students may be called upon (cold called) to answer questions. If you do not know the answer, you are permitted to pass.

EXAM DATES AND POLICIES

There is a take home mid-term exam and a closed book, in-person cumulative final exam in this course.

Midterm Exam: The midterm will be given the week of 10/19/22. Although it is a take home, your midterm must still be your own work without any assistance from others.

Final Exam: The final exam will be in-person at the time specified by the registrar. The syllabus will be updated to include the time after the registrar makes it available. Unless announced otherwise, the exam will be held in our assigned room for the term.

GRADING POLICY

Course grades are determined based on the following categories of work:

- **Class Attendance.** Attendance will be taken with Qwickly. Your attendance grade will be the percentage of class meetings you attend. Excused absences will not be counted toward your grade. Attendance is worth 3% of your grade.

- **Team Participation:** As described in the Classroom Conduct Section, you will be assigned to a team for learning collaboratively with your peers. Your contribution to your team counts for 5% of your grade.
- **Team Class Presentation:** As described in the Classroom Conduct Section, each team will also have the responsibility for presenting a set of papers for one of the classes. Teams will consult with me in advance of their presentation, and every member must take an active role in doing the presentation. This presentation is worth 5% of your grade.
- **Homework:** "Put it into practice" activities described in the timetable may have deliverables, and other exercises will be assigned as needed. This category is worth 5% of your final grade. Late homework will not be accepted.
- **Individual Project:** You are required to do an individual term project. Master's students may choose any of the following types of projects. PhD students are required to choose one of the first three types.

  - **Survey paper.** (Read at least 6 papers on the topic.)

    Use Google Scholar, ACM Portal and DBLP to find papers, focusing on those published in the following conferences: VLDB, SIGMOD, and ICDE. Depending on your topic, SIGOPS may also be appropriate. Feel free to see me for guidance on conference selection.

    Write a survey that includes an introduction, problem definition (including motivation and application domain), summary of techniques developed in each paper, global view of the papers covered, and future work suggestions. The length should be limited to and not exceed 6 pages in ACM conference format: https://www.acm.org/publications/proceedings-template

    You will be called to discuss your survey, and it will be evaluated on (a) understanding of the topic, (b) presentation and structure, and (c) critique of the research covered.

  - **Own research.**

    Proceed in the same manner as for the survey option above. In addition, identify a new research problem in the area and develop your own solution. Submit a paper describing your work. Your paper should include a motivation that shows how your work addresses a problem that related work did not address. It should compare your solution with related work. If your work includes experimental results, be sure to make a clear separation between the presentation of the measurements and your interpretation of them. You will be called to discuss your work. Your work will be evaluated for originality and novelty, and convincing argument or experimental results. In this case, the comprehensiveness of survey becomes secondary.

  - **Build a prototype.**

Identify a problem and examine existing solutions, using the instructions provided above. Implement one of the solutions, as found in a rank-1 conference (i.e., VLDB, SIGMOD, ICDE, SIGOPS) or premium journal paper (i.e., ACM TODS, VLDB Journal, IEEE TKDE, ACM TOCS). Feel free to see me for guidance on conference/paper selection. Write a 4-6 pages report, using ACM format as above. Include a discussion of the problem and the solution, and your experimental results. Try to reproduce some of the results in the paper. Submit the report along with a zip file of your code. Your report should explain whether you confirmed the published results or found some discrepancy, and what your result means. You will be called to demonstrate your prototype, and the work will be evaluated on (a) report quality and (b) demonstration effectiveness.

- **Master's Students Only: Build an application.**

  Identify an application of one the database systems related to the course content. Build an application of the database on AWS. Write a 4-6 pages report, using ACM format as above. Include a discussion of the problem your application solves and the solution. Discuss how your work illustrates, extends or diverges from the research in the area discussed in the course. Discuss what you learned and your suggestions for future work. Submit the report along with a zip file of your code. You will be called to demonstrate your application, and the work will be evaluated on (a) report quality and (b) demonstration effectiveness.

- Your project must be approved. To obtain approval, submit a proposal for your project by 10/1/2022.

  What if I'm late completing the Individual Project? If you are unprepared to discuss or demonstrate your work during the designated time at the end of term, you will lose the points for that part of the project grade. For the remainder, late submission of your work will be penalized as follows:

  - 1 day late, grace period with no points off
  - 2-3 days late, 3% off per day
  - 4^{th} day late, 4% off
  - 5-10 days late, 5% off per day
  - 11 or more, 10% off per day until no points are available and the grade is zero.

- **Final exam:** The final exam will be in person at the time specified by the registrar. It is closed-book, cumulative and worth 30% of your grade.

The following summarizes how each category of work contributes to your final numerical grade:

Class Attendance	3%
Team Participation	5%
Team Class Presentation	5%

Homework	5%
Midterm	22%
Individual Project	30%
Final Exam	30%

Grades will be assigned as follows from your final numeric grade:

A: 90-100	B+: 85-89	C+: 75-79	D: 60-69	F: 0-59
	B: 80-84	C: 70-74		

Other important notes:

- In addition to the ability to answer homework type problems, exams will also test your conceptual understanding of material, and your ability to apply it and extend it. Are you able to synthesize solutions to new problems from what you have learned? Are you able to solve problems related to the course creatively even if you have not previously seen them?
- There is NO extra credit. Plan to earn enough points to pass the course.

TENATIVE COURSE SCHEDULE

Wk.	Date	Topic	Notes
Introduction to Course and Cloud			
1	9/7	**Cloud**	
		While this is the first class and many are reluctant to start before that day, doing some of this reading before class will helpful. The following articles will familiarize you with cloud computing. Read them with the awareness that cloud computing is often hyped, and discussions of cloud computing can vary widely in emphasis since this area of computing is evolving rapidly. Goldman, D. What is the cloud? (2014) CNN. (2 pages). https://money.cnn.com/2014/09/03/technology/enterprise/what-is-the-cloud/index.html An excerpt from Lisdorf, A. (2021). "Introduction" in *Cloud Computing Basics: A Non-Technical Introduction*. Apres. (2 pages). Rutgers Library: https://link-springer-com.proxy.libraries.rutgers.edu/book/10.1007/978-1-4842-6921-3. How Cloud Computing Became a Big Tech Battleground. (2019). Wall Street Journal. (4 minutes, 16 seconds). https://www.youtube.com/watch?v=p7MqvJAKLoM	

Wk.	Date	Topic	Notes
		Mell, P., & Grance, T. (2011). Section 2 in The NIST definition of cloud computing. National Institute of Standards, Publication 800-145, pp. 2-3. (2 pages). https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf Ranger, S. What is cloud computing? Everything you need to know about cloud explained. (2022). ZDNet. (14 pages). https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-about-the-cloud/ Laberis, B. (2019). The disruptive force of cloud native. Natunix. (4 pages). https://www.nutanix.com/theforecastbynutanix/technology/the-disruptive-force-of-cloud-native While older, the following article is acknowledged as the first, best account of the differentiating features and issues in cloud computing. Some of the issues it mentions may have been fully addressed, but most are still issues today. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50-58. (9 pages) https://github.com/rxin/db-readings/blob/master/papers/cloud-computing.pdf	
2	**9/14**	**Cloud Architectures. Putting it together with AWS.**	
		<u>Put what we covered last time into practice:</u> Introduction, Modules 1-4 including the Knowledge Checks, and Lab 1, AWS Academy Cloud Foundations. **<u>Preparing for today's class:</u>** For IBM Cloud resources, feel free to skip IBM-specific product information. IBM Cloud Team (2021). Containers vs. virtual machines (VMs): What's the difference? IBM. (4 pages plus 13 minutes and 17 seconds of video). https://www.ibm.com/cloud/blog/containers-vs-vms IBM Cloud Education (2021). Docker. IBM. (7 pages plus 10 minutes 59 seconds of video). https://www.ibm.com/cloud/learn/docker IBM Cloud Education (2020). Continuous Integration. (7 pages). https://www.ibm.com/cloud/learn/continuous-integration IBM Cloud Education (2019). Continuous Deployment. (7 pages plus 13 minutes and 56 seconds of video). https://www.ibm.com/cloud/learn/continuous-deployment	

Wk.	Date	Topic	Notes
		Hoff, T. (2011). "Netflix: Developing, deploying, and supporting software according to the way of the cloud." Published in High scalability: Building bigger, faster, more reliable websites. (3 pages) http://highscalability.com/blog/2011/12/12/netflix-developing-deploying-and-supporting-software-accordi.html Bosch, J. (2015). Speed, data, and ecosystems: the future of software engineering. *IEEE Software*, *33*(1), 82-88. (6 pages). Available from the Rutgers Library: https://ieeexplore-ieee-org.proxy.libraries.rutgers.edu/stamp/stamp.jsp?tp=&arnumber=7368022 Savor, T., Douglas, M., Gentili, M., Williams, L., Beck, K., & Stumm, M. (2016, May). Continuous deployment at Facebook and OANDA. In *2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C)* (pp. 21-30). IEEE. (10 pages) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/2889160.2889223 Alary, H. (2018). "From bare-metal to Kubernetes." Published in Hugh Alary's blog. (8 pages) https://boxunix.com/2018/12/10/from-bare-metal-to-kubernetes/	
Introduction to the Big Data and the 4 V's: Volume, Variety, Velocity and Veracity			
3	**9/21**	**Big Data, Map/Reduce**	
		<u>Put what we covered last time into practice:</u> Modules 5-6 including the Knowledge Checks and Labs 2 and 3, AWS Academy Cloud Foundations. **<u>Preparing for today's class:</u>** Ellingwood, J. (2016). An Introduction to Big Data Concepts and Terminology. DigitalOcean. (6 pages) https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology Harrison, G. (2016). Chapter 2: Google, Big Data, and Hadoop. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 21-38. Apres. Read through the subsection on distributed relational databases only. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113. (7 pages) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/1327452.1327492 (In 2012, Dean	

Wk.	Date	Topic	Notes
		and Ghemawat, won the Association of Computing Machinery (ACM) Prize in Comuting for "their leadership in the science and engineering of Internet-scale distributed systems," including MapReduce.) For IBM Cloud resources, feel free to skip IBM-specific product information. IBM Cloud Education (2020). Data Warehouse. (9 pages plus 5 minutes and 17 seconds of video). https://www.ibm.com/cloud/learn/data-warehouse Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., ... & Murthy, R. (2010, March). Hive-a petabyte scale data warehouse using hadoop. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)* (pp. 996-1005). IEEE. (10 pages) https://ieeexplore-ieee-org.proxy.libraries.rutgers.edu/document/5447738 (The developers of Hive and Pig received the 2018 ACM SIGMOD Systems Award for their pioneering software systems that brought "relational-style declarative programming to the Hadoop ecosystem" which includes MapReduce. The paper describing Pig is in the recommended readings.) Recommended readings: Lin, J., & Dyer, C. (2010). Chapter 1: MapReduce basics. Published in Data-intensive text processing with MapReduce. *Synthesis Lectures on Human Language Technologies*, *3*(1), 18-38. https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008, June). Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1099-1110). Rutgers library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/1376616.1376726	
Addressing Volume			
4	**9/28**	**CAP, Scalability and Elasticity, Intro to Key-Value Databases with Amazon's DynamoDB**	
		<u>Put what we covered last time into practice:</u> Modules 7 with Knowledge Checks and Labs 4, AWS Academy Cloud Foundations. MapReduce Exercise and Hive Exercise in the AWS Learner Lab.	

Wk.	Date	Topic	Notes
		Preparing for today's class: Garcia-Molina, H., Ullman, J., & Widom, J. (2009). 20.3 Distributed Databases, 20.3.1 Distribution of Data, 2.3.2 Distributed Transactions, 2.3.3 Replication, 20.5 Distributed Commit (including subsections 20.5.1, 20.5.2, and 20.5.3). Published in *Database Systems: The Complete Book* (2nd ed.), pp. 997-999, 1008-1013. Pearson Education. (9 pages) Available from the Rutgers Library: https://bit.ly/3pqzHFq Carpenter, J. & Hewitt, E. (2016). Beyond relational databases. Published in *Cassandra: the definitive guide* (2nd ed.), 1-15. O'Reilly Media, Inc. (15 pages) Search the Internet for Business Applications of NoSQL Databases. See Canvas assignment for more details. Harrison, G. (2016). Chapter 3: Sharding, Amazon and the Birth of NoSQL. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 39-52. Apres. (14 pages) Abadi D. (2012). Consistency Tradeoffs in Modern Distributed Database System Design: CAP is Only Part of the Story. Computer (Long Beach, Calif). 45(2):37-42. doi:10.1109/MC.2012.33. (6 pages) https://ieeexplore-ieee-org.proxy.libraries.rutgers.edu/stamp/stamp.jsp?tp=&arnumber=6127847	
5	10/5	**Key-Value Database: Amazon's DynamoDB**	
		Put what we've covered into practice and extend that knowledge: Modules 8 with Knowledge Check and Lab 5, AWS Academy Cloud Foundations. **Do this exercise in the AWS Cloud Foundations Course Sandbox:** Perkins, L., Redmond, E., & Wilson, J. (2018). Chapter 7: DynamoDB. Published in *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf. Source code for examples is available at: https://pragprog.com/titles/pwrdata/seven-databases-in-seven-weeks-second-edition/ **Preparing for today's class:** DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., ... & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. Published in the Proceedings of the 2007 Symposium on Operating Systems (SOSP '07), *ACM SIGOPS operating systems review*, *41*(6), 205-220. (16 pages) https://dl.acm.org/doi/10.1145/1323293.1294281	

Wk.	Date	Topic	Notes
6	10/12	**Wide-Column Store: Google's BigTable and Facebook's Cassandra.**	
		<u>Preparing for today's class:</u> Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, *26*(2), 1-26. (27 pages) https://dl.acm.org/doi/10.1145/1365815.1365816 Carpenter, J. & Hewitt, E. (2022). Introducing Cassandra. Published in *Cassandra: the definitive guide* (2nd ed.), 16-33. O'Reilly Media, Inc. (27 pages) Lakshman, A., & Malik, P. (2010). Cassandra: a decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, *44*(2), 35-40. (6 pages) https://dl.acm.org/doi/10.1145/1773912.1773922 **Recommended reading. The second article is from Google on building a relational-style (NewSQL) database called Megastore on top of BigTable. Megastore powers Google's App Engine. If you skip Section 4 through 4.9, you can still get the gist. If you want to read Section 4, best to read the article about Paxos first.** Krzyzanowski, P. (2018). Understanding Paxos: Asynchronous Fault-Tolerant Consensus. (9 pages) https://people.cs.rutgers.edu/~pxk/417/notes/paxos.html Baker, Jason, Chris Bond, James C. Corbett, J. J. Furman, Andrey Khorlin, James Larson, Jean-Michel Leon, Yawei Li, Alexander Lloyd, and Vadim Yushprakh. (2011). Megastore: Providing scalable, highly available storage for interactive services. Published in the *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR '11)*, 223-234. (12 pages) (12 pages). https://storage.googleapis.com/pub-tools-public-publication-data/pdf/36971.pdf	
Addressing Variety			
7	10/19	**Document Stores and MongoDB**	
		<u>Put what we covered last time into practice:</u> **<u>Do the exercises in this chapter locally on your computer and then do the cloud part in the AWS Cloud Foundations Course Sandbox:</u>**	

Wk.	Date	Topic	Notes
		Perkins, L., Redmond, E., & Wilson, J. (2018). Chapter 3: HBase. Published in *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf. Source code for examples is available at: https://pragprog.com/titles/pwrdata/seven-databases-in-seven-weeks-second-edition/ **Preparing for today's class:** Harrison, G. (2016). Chapter 4: Document databases. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 53-64. Apres. Harrison, G. (2016). Chapter 8: Distributed database patterns. Published in *Next generation databases: NoSQL, newSQL, and big data*. Apres. Read the subsection on MongoDB only. Copeland, R. (2013). To Embed or Reference. Published in MongoDB Applied Design Patterns: Practical Use Cases with the Leading NoSQL Database, pp. 3-14. O'Reilly Media, Inc. (12 Pages) Available on reserve in the Rutgers Library. Note: MongoDB added transactions in Version 4.0 (2018) with enhancements in Version 4.2 (2019). Schultz, W., Avitabile, T., & Cabral, A. (2019). Tunable consistency in mongodb. *Proceedings of the VLDB Endowment*, *12*(12), 2071-2081. This is one of the few published research papers on MongoDB.	
8	**10/26**	**Graph Databases**	
		Put what we covered last time into practice: Perkins, L., Redmond, E., & Wilson, J. (2018). Chapter 4: MongoDB. Published in *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf. Source code for examples is available at: https://pragprog.com/titles/pwrdata/seven-databases-in-seven-weeks-second-edition/ Module 9 with Knowledge Checks, AWS Academy Cloud Foundations. Perform a three node MongoDB Deployment in the Cloud using resources from Chapter 4 of Perkins above and: Shukla, V., MongoDB on the AWS Cloud: Quick Start Reference Deployment (2018). https://aws.amazon.com/quickstart/architecture/mongodb/. **Preparing for today's class:**	

Wk.	Date	Topic	Notes
		Harrison, G.Harrison. (2016). Chapter 5: Tables are not your friends: Graph databases. Published in *Next generation databases: NoSQL, newSQL, and big data.*, pp.65-74. Apres. Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., & Czajkowski, G. (2010, June). Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 135-146. https://dl.acm.org/doi/10.1145/1807167.1807184 Check out the Stanford Network Analysis Project (SNAP) for some ideas about what can be represented in graphs and the results that can be obtained by analyzing them.	
9	**11/2**	**Integrating Big Data**	
		Put what we covered last time into practice: Perkins, L., Redmond, E., & Wilson, J. (2018). Chapter 6: Neo4J. Published in *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf. Source code for examples is available at: https://pragprog.com/titles/pwrdata/seven-databases-in-seven-weeks-second-edition/ **Preparing for today's class:** Dong, X. L., & Srivastava, D. (2015). Chapter 1: Motivation: Challenges and opportunities for BDI, and Chapter 2: Schema Alignment. Published in Big data integration. *Synthesis Lectures on Data Management*, *7*(1), 1-29.	
Addressing Velocity			
10	**11/9**	**Sources of Velocity. Streaming Systems.**	
		Put what we been covering into practice: Module 10 with Knowledge Check, AWS Academy Cloud Foundations. Congratulations, you have completed the AWS Course. **Preparing for today's class:** Lee, E. A., Hartmann, B., Kubiatowicz, J., Rosing, T. S., Wawrzynek, J., Wessel, D., ... & Rowe, A. (2014). The swarm at the edge of the cloud. *IEEE Design & Test*, *31*(3), 8-20.	

Wk.	Date	Topic	Notes
		Kleppmann, M. (2016). Chapter 1. Events and stream processing, and Chapter 2: Using logs to build a solid data infrastructure. Published in *Making sense of stream processing*, 1-79 (79 pages). O'Reilly Media, Inc. https://assets.confluent.io/m/2a60fabedb2dfbb1/original/20190307-EB-Making_Sense_of_Stream_Processing_Confluent.pdf Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R. J., Lax, R., ... & Whittle, S. (2015). The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. Published in *Proceedings of the VLDB Endowment* (Vol. 8), 1792-1803. https://research.google/pubs/pub43864/	
11	**11/16**	**Spark**	
		<u>Put what we covered last time into practice:</u> Damji, J., Lee, D., Wenig, B., & Das, T. (2020). Chapter 1: Introduction to Apache Spark: A unified analytics engine, and Chapter 2: Downloading Apache Spark and getting started. Published in *Learning spark: lightning-fast big data analysis* (2nd ed.), pp. 1-42. O'Reilly Media, Inc. (42 pages). Additional exercise to be determined. **<u>Preparing for today's class:</u>** Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56-65. https://dl-acm-org.proxy.libraries.rutgers.edu/doi/pdf/10.1145/2934664 Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... & Stoica, I. (2012). Resilient Distributed Datasets: A {Fault-Tolerant} Abstraction for {In-Memory} Cluster Computing. Published in the *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, 15-28. https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia	
12	**11/23**	**Change in Designation Day: No Class**	
13	**11/30**	**Spark Streaming**	
		<u>Preparing for today's class:</u>	

Wk.	Date	Topic	Notes
		Damji, J., Lee, D., Wenig, B., & Das, T. (2020). Chapter 8: Structured streaming. Published in *Learning spark: lightning-fast big data analysis* (2nd ed.), pp. 207-264. O'Reilly Media, Inc. (58 pages). Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., & Stoica, I. (2013, November). Discretized streams: Fault-tolerant streaming computation at scale. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles (SOSP)*, 423-438. https://dl.acm.org/doi/10.1145/2517349.2522737	
Addressing Veracity and Keeping Virtue			
14	**12/7**	**Veracity and Virtue**	
		Muniswamy-Reddy, K. K., Macko, P., & Seltzer, M. I. (2010, February). Provenance for the Cloud. Published in the *Proceedings of the File and Storage Technologies Conference (FAST)* (Vol. 10), 15-14. https://www.usenix.org/legacy/event/fast10/tech/full_papers/muniswamy-reddy.pdf Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2012). Truth finding on the deep web: Is the problem solved? Published in the *Proceedings of the VLDB Endowment*, 6(2), 97-108. Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., & Weikum, G. (2017, June). Fides: Towards a platform for responsible data science. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1-6. https://dl.acm.org/doi/10.1145/3085504.3085530 Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679. https://www.tandfonline.com/doi/full/10.1080/1369118X.2012.678878	
15	**12/14**	**Final Project Presentations**	

***Tentative schedule, subject to change. Check Canvas for the most up to date information on the schedule, readings and assignments.**

IMPORTANT DATES

Start of Classes:	Tuesday, September 6, 2022
End of Drop/Add Period:	Thursday, September 15, 2022
Thanksgiving Break:	Thursday, November 24 – Sunday, November 27
Last Day to Withdraw with W Grade:	TBD
End of Classes:	Wednesday, December 14, 2022
Exam Schedule:	See section above called “Exam Dates and Policies”