RUTGERS
Rutgers Business School
Newark and New Brunswick

Information Technology
Course Number: 22:544:645
Course Title: Advanced Database Systems

COURSE DESCRIPTION

This course focuses on research and applications in advanced database systems for Cloud and Big Data Computing. The course addresses each stage in the Data Science/Machine Learning (ML) Data Pipeline. It covers a variety of database architectures for big data: data warehouses, data lakes and data lakehouses. It illustrates these architectures in Amazon Web Services (AWS), Snowflake and Databricks.

This course provides an opportunity to learn about the end-to-end process of acquiring, preparing, storing, processing and using big data in data science. Students apply their learning on popular cloud platforms. The course topics include how to address the five V's of Big Data: volume, variety, velocity, veracity, and value. We also address how to maintain the virtue of our data, a sixth 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 and data processing techniques to address the challenges presented by big data. This course covers key-value stores, wide-column databases, document databases, graph databases and streaming data systems.

Key-value stores form the foundation for fast, incrementally scalable, distributed processing of Internet shopping carts, user information, and product information. We discuss Amazon's DynamoDB as an example of key-value stores. Wide-column databases support fast information storage and retrieval for search engines, personalization of services, analytics, and email. Google's BigTable and Facebook's Cassandra are our examples of wide-column databases.

Our example of a document database is MongoDB. MongoDB undergirds the high performance of many web sites and web applications. It is currently the most popular NoSQL database. Graph databases support analyzing social media relationships, transportation systems, and disease outbreaks. These databases increasingly find a role in automating machine learning pipelines, and are illustrated by Neo4j and Pregel. Data generated at high velocity such as data generated by sensors in the Internet of Things (IOT) require a streaming data system. We dive into these systems using Google Dataflow, Apache Beam, and Amazon Kinesis.

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 and Spark make these systems possible.

During the semester, 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. Free access to

Snowflake is provided through Snowflake for Academia.

In this course, class meetings will be a combination of lecture, discussion, team presentations and group exercises. Students will build and manipulate various databases and cloud services. Students will build applications on their own laptops and in the cloud.

Course Delivery Mode: All sections of this class are in-person.

Learning Management System: Canvas

Student Expectations:
In this course you will be expected to complete a number of tasks including:

- doing readings and assignments on time
- downloading and uploading documents and code to Canvas
- accessing online resources including cloud platforms, tools, articles, tutorials, and videos
- creating Flip videos
- communicating via GroupMe
- completing three synchronous quizzes using the Respondus Lockdown Browser
- participating in in-person class meetings and in team activities
- using an interactive tool on a cell phone or laptop to participate in class discussions
- making a presentation as part of a team
- completing an independent project

Hardware and software requirements:

- a webcam for recording a few Flip videos,
- a cell phone with web browser or a laptop in class to participate in slido interactions and discussions,
- a laptop in class for pre-announced working sessions or for use in taking exams, and

- at least 3 gigabytes of free space on a Windows laptop or 1 gigabyte on a Mac to install local tools. If you do not have the required space free, consider transferring some files, e.g. pictures and videos, to a usb drive or Google Drive until the end of the semester.

These requirements are satisfied by the minimum hardware recommended by OTIS:

- I5 Processor
- Windows 10 Professional
- 8gb of RAM
- 256gb hard drive (provided gigabytes are available for software installation)
- 720p webcam
- Internal mic

RBS Newark Students in need of financial assistance may submit their request via a form: https://myrun.newark.rutgers.edu/care-team

RBS New Brunswick Students in need of financial assistance can send an email to: deanofstudents@echo.rutgers.edu

Students can also benefit from reviewing: https://myrbs.business.rutgers.edu/students/learning-remotely

If students have any technology issues, they should reach out to OTIS help desk at helpdesk@business.rutgers.edu

Feedback and response expectations:

- If you have a question that you think is shared by your classmates, please ask the question on your section's GroupMe. I will answer your question there, so everyone will learn the answer.
- A GroupMe Direct Message (DM) or email to my Scarlet Mail address are the best ways to reach me about our course. Unlike Canvas Direct Messages, GroupMe DM and Scarlet Mail provide context for continuing conversations. Emailing might result in a response delay because it does not notify my cell phone. GroupMe and Scarlet Mail do notify my cell phone.
- Set up Canvas to notify you of course announcements. Check Canvas often. You are responsible for meeting the deadlines and fulfilling the requirements posted on Canvas.
- Email/Direct Messaging Response Times: I will do my best to respond to your emails quickly, often within hours, but at a maximum within 24 hours weekdays and 48 hours on weekends. During Break, email responses may be delayed until class resumes. Please remind me if you do not hear back from me within this time.
- Graded Materials Return Times: Many of your assignments will be automatically graded by AWS Academy. Similarly, most of your exam questions will be graded automatically by Canvas. Since I have classes with several large sections, I will strive to have your other assignments and test questions graded within two weeks, but three weeks may be required.

COURSE MATERIALS

- Required books:
  - Akidau, T., Chernyak, S., & Lax, R. (2018). *Streaming systems: the what, where, when, and how of large-scale data processing*. O'Reilly Media, Inc. Available through the library.
  - Carpenter, J. & Hewitt, E. (2022). *Cassandra: the definitive guide* (Revised 3rd ed.). O'Reilly Media, Inc. Available through the library.
  - Harrison, G. (2016). *Next generation databases: NoSQL, newSQL, and big data*. Apres. Available through the library.
  - Perkins, L., Redmond, E., & Wilson, J. (2018). *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf. Available through the library.
- Recommended books:
  - Damji, J., Lee, D., Wenig, B., & Das, T. (2020). *Learning Spark: lightning-fast big data analytics* (2nd ed.) O'Reilly Media, Inc. This book may be available free from some companies.
  - 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 schedule below.
  - Check Canvas (https://canvas.rutgers.edu/) and your Scarlet Mail Rutgers email account regularly for additional course materials. materials.

LEARNING GOALS AND OBJECTIVES

For Master of Information Technology and Analytics (MITA) Program:

This course satisfies the following Rutgers Business School goals and objectives:

- Business technology knowledge. Students graduating with a Master of Information Technology degree will be able to demonstrate business technology knowledge.

  Students will demonstrate:

  - Understanding of the current practices and technology used in businesses.
  - Ability to analyze and solve complex business problems with cutting edge technology.

- Information technology knowledge. Students graduating with a Master of Information Technology degree will be able to demonstrate information technology knowledge.

  Students will demonstrate:

  - Understanding of basic information technology concepts.
  - Ability to analyze and solve information technology problems.

- Critical thinking skills. Students graduating with a Master of Information Technology degree will be able to understand complex business situations and provide solutions to improve current business practices.

  Students will demonstrate:

  - Ability to identify problems in a situation.
  - Ability to find innovative solutions.

- Communication skills. Students graduating with a Master of Information Technology degree will be able to effectively communicate in a way that demonstrates sensitivity to an audience's needs.

  Students will demonstrate:

  - Ability to communicate information in a clear concise manner.
  - Ability to communicate relatively complex ideas in an understandable manner.

For Ph.D. in Management Program:

This course satisfies the following Rutgers Business School goals and objectives:

- Advanced Knowledge in Specialized Areas. Doctoral students will acquire advanced knowledge in areas of specialization.

- Advanced Research Skills. Doctoral students will develop advanced theoretical or practical research skills for the area of specialization.

PREREQUISITES

Students taking this course should have knowledge of relational database systems, including database normalization, entity relationship diagrams and design, and SQL, 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.

Use of AI such as ChatGPT is only permitted to help you brainstorm ideas and see examples. All material you submit for assignments must be your own. Use of AI such as ChatGPT is not permitted during exams.

Be aware that AI software, such as ChatGPT, may give a false answer to an assignment, or a truthful statement that does not satisfy the requirements of the assignment. Using either of these types of answers would lose full or partial credit on the assignment. You are responsible for the correctness and appropriateness of your answers, not ChatGPT.

CLASSROOM CONDUCT

Research has shown that students learn better in a community with their peers. I 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 an article, paper or technology during one of the class meetings. 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 discussion sessions. A series of students may be called upon (cold called) to answer questions or contribute an opinion. If what you know is insufficient to answer, you are permitted to pass.

EXAM DATES AND POLICIES

There are three quizzes in this course, but no exams. The quizzes are closed book and in-person. The quizzes focus on the most recent material, but the material builds through the term and is in that sense cumulative. The quizzes occur approximately every four weeks.

During the quizzes, the following rules apply:

- All tests are in-person.
- If you take an in-person test on Canvas while not present in person, you will be reported for a violation of academic integrity and receive a zero for the test.
- In the instructions for each test, the test will explicitly list the permitted material that you can use when taking the test. If you access any material that is not permitted, you will be reported for a violation of academic integrity and receive a zero for the test.
- If you qualify for accommodations that will influence testing procedures, have the Office of Disability Services send me a letter at the start of the semester. Reach out to discuss the accommodation with me.
- Cell phones must be turned off and placed in front of you facedown during the test.

- Use the bathroom prior to the start of the test.
- Your test will receive a grade of zero if you do not sign the Honor Pledge.

GRADING POLICY

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

- **Class Attendance.** Attendance will be taken with Qwickly. If you attend less than 75% of the class meeting on a particular day, you will not receive credit for attending on that day. Your attendance grade will be the percentage of class meetings you attend. Excused absences will not be counted toward your grade. Attendance is worth 2% 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 2% of your grade.
- **Team Class Presentation:** As described in the Classroom Conduct Section, each team will also have the responsibility for presenting a paper during one of the class meetings. 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 6% 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 30% of your final grade. Late homework may not be accepted.
- **Individual Project:** You are required to do an individual term project. The project is worth 30% of your grade. 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, other conferences such as SOSP or CIDR 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 to address the problem, 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 (references can be on additional pages): https://www.acm.org/publications/proceedings-template

      You will present your work, 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 present 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 one conference (i.e., VLDB, SIGMOD, ICDE, SOSP) 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 (references can be on additional pages) 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. You may use artifacts provided by the authors of the paper as part of your evaluation. Submit the report along with a zip file of your code and any code used from the authors of the paper. Include instructions for installing and running your system/solution in a file called README in your zip submission. Your report should explain whether you confirmed the published results or found some discrepancy, and what your results mean. You will present and 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 the database systems and data processing techniques related to the course content. Build an application of the system/techniques. Write a 4-6 pages report (references can be on additional pages) 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. Building and evaluating an application using two different systems/techniques from the course, and then comparing the results and your experiences can lead to a very compelling report. Submit the report along with a zip file of your code. Include instructions for installing and running your system/application in a file called README in your zip submission. 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.

  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
  - 4th 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.

- **Quizzes:** There are three in person quizzes. They are closed-book and worth 30% of your grade.

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

Class Attendance	2%
Team Participation	2%
Team Class Presentation	6%
Homework	30%
Quizzes	30%
Individual Project	30%

Grades will be assigned as follows from your final numeric grade for students in the master's degree section of the course (22:544:645:01):

	B+: 87-89	C+: 77-79	
A: 93-100	B: 83-86	C: 73-76	F: 0-69
A-: 90-92	B-: 80-82	C-: 70-72	

Fewer grades are available for students taking the PhD degree section of the course (26:198:641:01). Those grades are A, B+, B, C+, C and F. With respect to the scale given above, PhD students will receive the next lowest grade if the grade listed above is not available.

Other important notes:

- In addition to the ability to answer homework type problems, quizzes will also test your conceptual understanding of material, your ability to integrate what you have learned and analyze it, and ability to apply what you learned and extend it. Are you able discuss the tradeoffs between different strategies for addressing database consistency? Are you able

to suggest new approaches to solving database problems? Can you advise a company on big data strategies?

- There is NO extra credit. Plan to earn enough points to pass the course.

TENATIVE COURSE SCHEDULE

Wk.	Topic	Notes
1	**Cloud**	Preparation: 32 pages of reading and 00:18:39 of video.
	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. An excerpt from Lisdorf, A. (2021). "Introduction" in *Cloud Computing Basics: A Non-Technical Introduction*. Apres, pp. xiii-xv (3 pages). Available from the 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). 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). Ranger, S. What is cloud computing? Everything you need to know about cloud explained. (2022). ZDNet. (14 pages). How Cloud Giants Amazon, Google, and Microsoft Got Even Bigger. (2022). Wall Street Journal Tech News Briefing. (8 minutes, 6 seconds) Cloud Computing Isn't as Cost Effective as Hoped. So What's Next? (2022). Wall Street Journal Tech News Briefing. (6 minutes, 17 seconds) Laberis, B. (2019). The disruptive force of cloud native. Natunix. (4 pages).	
Wk.	**Topic**	**Notes**

	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) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/10.1145/1721654.1721672	
2	**Cloud Architectures. Putting it together with AWS.**	Preparation: 31 pages of reading + 01:07:38 of video. Team Presentation: 3-13 pages of reading depending on team assignment. Practice: AWS Modules: 02:15:35 of video, 2 pages of reading, one lab, and Knowledge Checks.
	Put what we covered last time into practice: Introduction, AWS Academy Cloud Foundations Modules 1-4 including Lab 1 and Knowledge Checks. **First preliminaries for database assignments:** See Canvas for an assignment to install tools. **Preparing for today's class:** 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 19 seconds of video). IBM Cloud Education (2021). Docker. IBM. (12 pages plus 10 minutes and 59 seconds of video). IBM Cloud Education (2020). Continuous Integration. (8 pages plus 6 minutes and 20 seconds of video). IBM Cloud Education (2019). Continuous Deployment. (7 pages plus 7 minutes and 36 seconds of video). **Team Presentations in Class:** 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)	
Wk.	**Topic**	**Notes**

		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 Watch the video of Tony Savor presenting the paper at (29 minutes and 26 seconds): https://www.youtube.com/watch?v=ERJZAwkHpX0 Alary, H. (2018). "From bare-metal to Kubernetes." Published in Hugh Alary's blog. (8 pages) Liguori, C. (2020). Automating safe, hands-off deployments. Amazon. (13 pages). **Recommended Article:** IBM Cloud Education (2019). What is Kubernetes? (14 pages plus 11 minutes and 57 seconds of new video. One of the videos is also in the Docker reference above.).	
Introduction to the Big Data and the 6 V's: Volume, Variety, Velocity, Veracity, Value & Virtue			
3		**Big Data, Data Warehouses, Data Lakes, and Data Pipelines**	Preparation: 58 pages reading. Practice: AWS Modules: approximately 01:59:39 of video, two labs, activities, and Knowledge Checks.
		Put what we covered last time into practice: AWS Academy Cloud Foundations Modules 5-6 including Labs 2 and 3, Activities, and Knowledge Checks. **Second preliminaries for database assignments:** See Canvas for an assignment on the Unix/Linux shell. **Preparing for today's class:** Ellingwood, J. (2016). An Introduction to Big Data Concepts and Terminology. DigitalOcean. (6 pages) Han J, Kamber M, Pei J. (2012). Sections through 4.4 from "Chapter 4: Data Warehousing and Online Analytical Processing" in *Data Mining Concepts and Techniques* (3rd ed.), pp. 125-165. Elsevier. (41 pages) Available from the Rutgers Library: https://bit.ly/44D4Tn9	

Wk.	Topic	Notes

	Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*, *12*(12), 1986-1989. (4 pages) Available from the Rutgers Library: https://bit.ly/3R5sCJu Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (Vol. 8). (7 pages) **Team Presentations in Class:** Overview/Demonstration of Snowflake. Overview/Demonstration of Databricks. **Recommended Articles:** Dageville, B., Cruanes, T., Zukowski, M., Antonov, V., Avanes, A., Bock, J., et al. (2016, June). The snowflake elastic data warehouse. In Proceedings of the 2016 International Conference on Management of Data (pp. 215-226). Available from the Rutgers Library: https://bit.ly/4dV7Y6R. Armenatzoglou, N., Basu, S., Bhanoori, N., Cai, M., Chainani, N., Chinta, K., et al. (2022, June). Amazon Redshift re-invented. In *Proceedings of the 2022 International Conference on Management of Data* (pp. 2205-2217). Available from the Rutgers Library: https://bit.ly/3XgSKns.
Addressing Volume	
4	**Big Data Processing, MapReduce** — Preparation: 34 pages reading. Practice: AWS Modules: approximately 02:16:02 of video, three labs, and Knowledge Checks. **Put what we covered last time into practice:** AWS Academy Cloud Foundations Module 7 including Lab 4 with Knowledge Check, and AWS Academy Data Engineering Modules 1-3, and 8 including the lab in Module 2 and the lab in Module 8 with Knowledge Checks. Module 1 is missing its video, so just flip through the slides. **Preparing for today's class:**
Wk.	**Topic** — **Notes**

		Harrison, G. (2016). Chapter 2: Google, Big Data, and Hadoop. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 21-37. Apres. (17 pages) 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 and Ghemawat, won the Association of Computing Machinery (ACM) Prize in Computing for "their leadership in the science and engineering of Internet-scale distributed systems," including MapReduce.) **Recommended Article:** 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
5	**Spark**	Preparation: 19 pages of reading + initial review of Spark Jupyter Notebooks. Practice: Map Reduce Exercise. **Quiz.**
	Put what we covered last time into practice: See Canvas for MapReduce assignment. **Preparing for today's class:** 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. (10 pages) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/pdf/10.1145/2934664 (In 2022, many contributors to Apache Spark won the ACM SIGMOD Systems Award for creating "an innovative, widely-used, open-source, unified data processing system encompassing relational, streaming, and machine-learning workloads." See https://sigmod.org/2022-sigmod-awards/ for more detail.) Spark on Google Colab (single node) Spark on AWS EMR Cluster (multiple node cluster)	
6	**Hive, Distributed Databases, CAP, Scalability and Elasticity**	Preparation: 49 pages of reading. Practice: AWS Modules: approximately 00:42:22 of video, two labs, Knowledge Checks, and Spark Exercise.
	Put what we covered last time into practice:	

Wk.	Topic	Notes

		AWS Academy Data Engineering Module 9 with two labs and Knowledge Checks. See Canvas for Spark assignment. **Preparing for today's class:** Garcia-Molina, H., Ullman, J., & Widom, J. (2009). 20.1 Parallel Algorithms on Relations. Published in *Database Systems: The Complete Book* (2nd ed.), pp. 985-993, 1008-1013. Pearson Education. (8 pages) Available from the Rutgers Library: https://bit.ly/3pqzHFq 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) Available from the Rutgers Library: https://ieeexplore-ieee-org.proxy.libraries.rutgers.edu/document/5447738 (The developers of Hive and Pig received the 2018 ACM Special Interest Group on Management of Data (SIGMOD) Systems Award for their pioneering software systems that brought "relational-style declarative programming to the Hadoop ecosystem." The Hadoop ecosystem includes MapReduce. The paper describing Pig is in the recommended readings.) 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. (2022). Beyond relational databases. Published in *Cassandra: the definitive guide* (Revised 3rd ed.), 1-16. O'Reilly Media, Inc. (16 pages) Available from the Rutgers Library: https://bit.ly/3z2A94W 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) Available from the Rutgers Library: https://ieeexplore-ieee-org.proxy.libraries.rutgers.edu/stamp/stamp.jsp?tp=&arnumber=6127847 (In 2009, Eric Brewer, the creator of the CAP Theorem, won the ACM Prize in Computing "for his design and development of highly scalable internet services and innovations in bringing information technology to developing regions.") **Recommended Article:**

Wk.	Topic	Notes

		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). Available from the Rutgers library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/1376616.1376726
7	**Intro to Key-Value Databases with Amazon's Dynamo. Intro to Wide- Column Databases with Google's BigTable.**	Preparation: 64 pages of reading. Practice: AWS Modules: approximately 01:18:37 of video, one lab, and Knowledge Checks.
	Put what we covered last time into practice: Data Engineering Modules 4-5 including the lab in Module 4, and Knowledge Checks. Note that the "Cloud security review" recording in Module 5 overlaps with Module 4 in the AWS Academy Cloud Foundations Course. **Preparing for today's class:** Harrison, G. (2016). Chapter 3: Sharding, Amazon and the Birth of NoSQL. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 39-51. Apres. (13 pages) Available from the Rutgers Library: https://bit.ly/4dQHN18 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) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/1323293.1294281 (In 2017, this paper received the ACM Special Interest Group on Operating Systems (SIGOPS) Hall of Fame award as the most influential paper over the previous decade.) Krzyzanowski, P. (2021). BigTable: A NoSQL wide-column single-table database. (8 pages) 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) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/1365815.1365816 (In 2016, the conference version of this paper won the ACM SIGOPS Hall of Fame award as the most influential paper published over the previous decade.)	
Wk.	**Topic**	**Notes**

	Recommended Readings. 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) 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).	
8	**Cassandra, introduced by Facebook in 2007, combining Wide-Column and Key-Value Database Features. Extended Microservice Example using Cassandra.**	Preparation: 67 pages of reading. Practice: AWS Modules: approximately 00:34:56 of video, one lab, and Knowledge Check.
	Put what we been covering into practice: AWS Academy Cloud Foundations Module 8 with Knowledge Check and Lab 5. **Preparing for today's class:** Carpenter, J. & Hewitt, E. (2022). Chapter 2: Introducing Cassandra. Published in *Cassandra: the definitive guide* (Revised 3rd ed.), 17-31. O'Reilly Media, Inc. (15 pages) Available from the Rutgers Library: https://bit.ly/4cCEeKY Carpenter, J. & Hewitt, E. (2022). Chapter 4: The Cassandra Query Language and Chapter 5: Data Modeling in Introducing Cassandra. Published in *Cassandra: the definitive guide* (Revised 3rd ed.), 55-106. O'Reilly Media, Inc. (52 pages) Available from the Rutgers Library: https://bit.ly/3Mk0pei and https://bit.ly/46YagQl **Team Presentation:** Demonstration of Cassandra using Amazon Keyspaces.	

Wk.	Topic	Notes

Addressing Variety		
9	**Document Stores and MongoDB**	42 pages of reading. **Quiz.**
	Preparing for today's class: Harrison, G. (2016). Chapter 4: Document databases. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 53-63. Apres. (11 pages) Available from the Rutgers Library: https://bit.ly/3Z2hT6F Harrison, G. (2016). Chapter 8: Distributed database patterns. Published in *Next generation databases: NoSQL, newSQL, and big data*, pp. 110-115. Apres. (5 pages) Read the subsection on MongoDB Sharding and Replication only. Available from the Rutgers Library: https://bit.ly/3yQFWuo Harrison, G. (2016). Chapter 11: Languages and programming interfaces. Published in *Next generation databases: NoSQL, newSQL, and big data*. pp. 173-175. Apres. (3 pages) Read the subsection on MongoDB only. Available from the Rutgers Library: https://bit.ly/4dQIEim 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. (11 pages) This is one of the few published research papers by MongoDB. Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.14778/3352063.3352125	
10	**Graph Databases, Integrating Big Data**	Preparation: 22 pages of reading. Practice: AWS Modules: 00:50:04 of video, one lab, and Knowledge Checks. MongoDB Exercise based on 40 pages of reading due next week.
	Put what we've been covering into practice: AWS Academy Cloud Foundations Modules 9 and 10 with Knowledge Checks and Lab 6. Complete the course assessment. You will receive a badge from AWS Academy. Congratulations, you have completed the AWS Academy Cloud Foundations Course.	

Wk.	Topic	Notes

	See Canvas for MongoDB Assignment due next week. The exercise is based on the following: 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*, pp. 93-133. Pragmatic Bookshelf. (40 pages) **Preparing for today's class:** Harrison, G. (2016). Chapter 5: Tables are not your friends: Graph databases. Published in *Next generation databases: NoSQL, newSQL, and big data*., pp.65-74. Apres. (10 pages) Available from the Rutgers Library: https://bit.ly/3z2BPeK 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. (12 pages) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/1807167.1807184 (In 2020, this paper received the ACM SIGMOD Test of Time Award as the most influential paper over the previous decade. The authors of this paper were recognized for the far-reaching influence of their work on graph databases and graph analytics.) 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.	
Addressing Velocity		
11	**Sources of Velocity. Streaming Systems.**	Preparation: 88 pages of reading. (Sounds like a lot, but the longest reading has lots of pictures and small pages.) Practice: MongoDB Exercise due this week. Neo4J Exercise based on 33 pages of reading due the week after Thanksgiving.
	Put what we've been covering into practice: MongoDB Assignment due this week. See Canvas for Neo4j Assignment due the week after Thanksgiving. 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*, pp. 177-209. Pragmatic Bookshelf. (33 pages) **Preparing for today's class:**	
Wk.	**Topic**	**Notes**

	Kleppmann, M. (2016). Chapter 1. Events and stream processing. Published in *Making sense of stream processing*, 1-37. (38 pages) O'Reilly Media, Inc. 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. (12 pages) Akidau, T., Chernyak, S., & Lax, R. (2018). "Chapter 2: The What, Where, When, and How of Data Processing" in *Streaming systems: the what, where, when, and how of large-scale data processing*, pp. 25-57. (33 pages) O'Reilly Media, Inc. Available through the library: https://bit.ly/3T0c6KU. Read with the online animated figures in: http://www.streamingbook.net/figures **Team Presentation:** Spark Streaming Demonstration.	
Addressing Veracity and Keeping Virtue, More on Data Pipelines		
12	**Veracity and Virtue, Data Ingestion and Preparation in the Pipeline**	Preparation: 29 pages of reading, and 01:32:05 of video, and 2 labs, and Knowledge Checks. Practice: Neo4j Exercise due next week. AWS exercises included in preparation assignment.
	Preparing for today's class: 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. (6 pages) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/abs/10.1145/3085504.3085530 Werder, K., Ramesh, B., & Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, *13*(2), 1-23. (23 pages) Available from the Rutgers Library: https://bit.ly/3R5JH67 AWS Academy Data Engineering Modules 6-8 on Data Ingestion and Preparation with two labs as you encounter them and Knowledge Checks.	
Wk.	**Topic**	**Notes**

12	**Change in Designation Day: No Class**	**Break**
13	**Final Quiz and Project Q and A**	**Quiz.**
	Put what we been covering into practice: Neo4j Assignment due this week. **Preparing for today's class:** Bring your project questions for discussion after the quiz.	
14	**Early Short Project Presentations. Data Pipelines and ML, Data Analysis and Visualization. Automated Pipelines.**	Preparation: 6 pages of reading, 02:15:37 of video, two labs, and Knowledge Checks.
	Note: Depending on the number of early presentations, the topics for this class meeting may change. **Preparing for today's class:** AWS Academy Data Engineering Modules 10-12 including the Module 11 Lab and Module 12 Lab, and Knowledge Checks. Complete the course assessment. You will receive a badge from AWS Academy. The Data Engineering Course is partial preparation for advanced AWS certifications. Check out the slides for Module 13 for more information. Schaefer, K. (2023). How autonomous agents will support data exploration, Presentation at the *3rd Google Women in Machine Learning Symposium*. The code assistant is now available for free in Colab. Amazon also provides a free code assistant, called Code Whisperer. Leskovec, J. (2023, June). Databases as Graphs: Predictive Queries for Declarative Machine Learning. In *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (pp. 1-1). (1 page) Available from the Rutgers Library: https://dl-acm-org.proxy.libraries.rutgers.edu/doi/10.1145/3584372.3589939	
Wk.	**Topic**	**Notes**

	Leskovec, J. (2023). Graphs, Databases and Machine Learning. Innovation Award Presentation at the *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Watch the recording at: https://www.youtube.com/watch?app=desktop&v=lmPMTW1Y2JM (At this conference, Jure Leskovec received the KDD Innovation Award which "recognizes individuals for their outstanding technical contributions to the field of knowledge discovery in data and data mining that have had lasting impact in furthering the theory and/or development of commercial systems." Kumo is a startup recently cofounded by Jure Leskovec.) How AI Accelerates ML Development: Kumo's approach to intelligent data science. (2024). Kumo. (5 pages) **Early Project Presentations**	
16	**Mini-Conference with Short Project Presentations**	

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

RUTGERS BELOVED COMMUNITY

The Rutgers Beloved Community is "defined by a commitment to work together to embody, reflect, and respect the complexity of all our parts." Bias incidents are contrary to the respect we show to each other.

Bias Incidents: A bias incident is an act – either verbal, written, physical, or psychological that threatens or harms a person or group on the basis of actual or perceived race, religion, color, sex, age, sexual orientation, gender identity or expression, national origin, ancestry, disability, marital

status, civil union status, domestic partnership status, atypical heredity or cellular blood trait, military service or veteran status.

Bias incidents can be reported online at:

New Brunswick campus - Bias Incident Report Form

Newark campus - Bias Incident Report Form

SUPPORT SERVICES

If you need accommodation for a ***disability***, obtain a Letter of Accommodation from the Office of Disability Services. The Office of Disability Services at Rutgers, The State University of New Jersey, provides student-centered and student-inclusive programming in compliance with the Americans with Disabilities Act of 1990, the Americans with Disabilities Act Amendments of 2008, Section 504 of the Rehabilitation Act of 1973, Section 508 of the Rehabilitation Act of 1998, and the New Jersey Law Against Discrimination. More information can be found at ods.rutgers.edu. Rutgers University-New Brunswick ODS phone (848)445-6800 or email dsoffice@echo.rutgers.edu

If you are ***pregnant***, the Office of Title IX and ADA Compliance is available to assist with any concerns or potential accommodations related to pregnancy. Rutgers University-New Brunswick Title IX Coordinator phone (848)932-8200 or email jackie.moran@rutgers.edu]

If you seek ***religious accommodations***, the Office of the Dean of Students is available to verify absences for religious observance, as needed. Rutgers University-New Brunswick Dean of Students phone (848)932-2300 or email deanofstudents@echo.rutgers.edu

If you have experienced any form of ***gender or sex-based discrimination or harassment***, including sexual assault, sexual harassment, relationship violence, or stalking, the Office for Violence Prevention and Victim Assistance provides help and support. More information can be found at http://vpva.rutgers.edu/. At Rutgers University-New Brunswick the incident report link is http://studentconduct.rutgers.edu/concern/. If you wish to speak with a staff member who is confidential and does **not** have a reporting responsibility, you may contact the Office for Violence Prevention and Victim Assistance at (848) 932-1181.

If you have experienced a temporary condition or injury that is adversely affecting your ability to fully participate, you should submit a request via https://temporaryconditions.rutgers.edu .

If you are a military ***veteran*** or are on active military duty, you can obtain support through the Office of Veteran and Military Programs and Services. http://veterans.rutgers.edu/

If you are in need of ***mental health*** services, please use our readily available services. Rutgers Counseling and Psychological Services–New Brunswick: http://rhscaps.rutgers.edu/.

If you are in need of ***physical health*** services, please use our readily available services. Rutgers Health Services – New Brunswick: http://health.rutgers.edu/.

If you are in need of ***legal*** services, please use our readily available services: http://rusls.rutgers.edu/

Students experiencing difficulty in courses due to ***English as a second language (ESL)*** should contact the Program in American Language Studies for supports. Rutgers–New Brunswick: eslpals@english.rutgers.edu.

If you are in need of additional ***academic assistance***, please use our readily available services. Rutgers University-New Brunswick Learning Center: https://rlc.rutgers.edu/.

CODE OF PROFESSIONAL CONDUCT

(This code of conduct is also available at: https://myrbs.business.rutgers.edu/students/code-professional-conduct.)

Rutgers Business School is recognized for its high-quality education. To that end, maintaining the caliber of classroom excellence requires students to adhere to the same behaviors that are expected in professional career environments. These include the following principles:

Discussion and Correspondence

- Each student is encouraged to take an active part in class discussions and activities. Substantive dialogue requires a degree of mutual respect, willingness to listen, and tolerance of opposing points of view. Disagreement and the challenging of ideas must happen in a supportive and sensitive manner. Hostility and disrespectful behavior will not be tolerated.
- In both correspondence and the classroom, students should demonstrate respect in the way they address instructors. Students should use proper titles in addressing instructors unless there is an explicit understanding that the instructor accepts less formal address. Similarly, appropriate formatting in electronic communication, as well as timely responsiveness, are all expectations in every professional interaction, including with instructors. Everything said and written should demonstrate respect and goodwill.

Punctuality and Disruption

- Class starts and ends promptly at the assigned periods. Students are expected to be in their seats and ready to begin class on time.
- Packing belongings before the end of class is disruptive to both other students and the instructor. Barring emergencies and within reason, students are expected to remain in their seats for the duration of the class.

Technology

- The use of technology is sanctioned only as permitted by the course instructor. As research on learning shows, peripheral use of technology in classes negatively impacts the learning environment in three ways:

    1. Individual learning and performance directly suffer, resulting in the systemic lowering of grades earned.

    2. One student's use of technology automatically diverts and captures other people's attention, thus impeding their learning and performance. Moreover, even minor infractions have a spillover effect and result in others doing the same.

    3. Subverting this policy (e.g., using a phone during class, even if hidden below the table; tapping on a smartwatch; using a laptop for non-course related matters) is evident to the course instructor and offensive to the principles of decorum in a learning environment.

- Networking, computing, and associated resources in the trading rooms, advanced technology rooms, and general classrooms are to be used in the manner intended.
- Sharing links to private online classes, attempting to join an online class that you are not enrolled in, or posting disruptive content during these sessions are strictly prohibited and may lead to disciplinary action.
- For more instructions on information technology resources at Rutgers University, please refer to the Acceptable Use Policy for Information Technology Resources.

Misappropriating Intellectual Property

- Almost all original work that is available to you is subject to claims of copyright by its creators or copyright holders. These copyright holders may include publishers, authors, professors, the University, RBS, and in some cases, your fellow students. The protected materials may include but are not limited to syllabi, recorded lectures, PowerPoint presentations, and other recorded, printed, or electronically stored media. These materials are only limited to completing the requirements of the class.

- Unauthorized use includes such things as copying, sharing, forwarding, selling, renting, online posting, publication, or any other form of distribution of these materials without the written permission of the copyright holder. Such misconduct may potentially subject you to disciplinary action by the University, significant civil penalties, and even severe criminal sanctions.

- For more instructions on copyright protections at Rutgers University, please refer to the Rutgers Libraries.

Rutgers Business School is committed to the highest standards of integrity. We value mutual respect and responsibility, as these are fundamental to our educational excellence both inside and outside the classroom.