SYLLABUS

Big Data Technologies

Instructor

Yousef Elmehdwi, Associate Teaching Professor of Computer Science, Illinois Institute of Technology
Yousef Elmehdwi | Illinois Institute of Technology

Course Description

Big data is the area of informatics focusing on datasets whose size is beyond the ability of typical database and other software tools to capture, store, analyze and manage. This course provides a rapid immersion into the area of big data and the technologies which have recently emerged to manage it.

We start with an introduction to the characteristics of big data and an overview of the associated technology landscape and continue with an in depth exploration of Hadoop, the leading open source framework for big data processing. Here the focus is on the most important Hadoop components such as Hive, Pig, stream processing and Spark as well as architectural patterns for applying these components. We continue with an exploration of the range of specialized (NoSQL) database systems architected to address the challenges of managing large volumes of data.

Overall the objective is to develop a sense of how to make sound decisions in the adoption and use of these technologies as well as economically deploy them on modern cloud computing infrastructure.

Course Outcomes

Upon successful completion of this course, you will be able to:

- Understand the history and trends of the data analytics field via key industry development projects and academic research topics
- Understand the core cluster architecture and database management system technologies used in big data systems
- Study the key trade-offs between scaling of vertically integrated systems versus horizontally sharded systems
- Comprehend reliability and fault-tolerance in distributed operating systems versus traditional operating systems in the context of storage/compute clusters (File Systems & Task Schedulers)
- Study the key trade-offs between NoSQL distributed database management systems versus traditional relational database management systems
- Comprehend availability and consistency in distributed databases versus traditional databases in the context of node consensus/network partitions (Paxos/Raft, CAP Theorem)
- Develop familiarity with the the Apache software platform (Hadoop [HDFS/YARN], Hive, Pig, Spark, etc.) environment for developing and operating big-data model pipelines.
- Develop familiarity with AWS cloud infrastructure (S3, EC2, EMR, etc.) capabilities for designing and architecting big-data system instances.
- Learn from various open access data/articles for research papers and analytical work.
- Build with various open source software/code for development projects and logistical work.

Course Materials

The link to reading materials and resources to learn on the topics can be found in each week's learning module. All materials are available online for free, no required resources need to be purchased. Note: Be aware that some resources may open in a new tab.

Software Requirements:

- Access to a Linux virtual machine or Unix-like operating system
- Python environment/virtual-environment
- AWS account and API access

Course Outline

The course consists of 8 modules that focus on the following key areas:

Module 1: Big Data Concepts

Key concepts

- From Data to Value
- Big Data Overview
- Confounding Factors
- Big Data Challenges
- Big Data Benefits
- Big Data Technology
- Generic Distributed Storage Systems and Execution Engines

Readings

- Sarker, I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN COMPUT. SCI. 2, 377 (2021). https://doi.org/10.1007/s42979-021-00765-8
- H. Hu, Y. Wen, T. -S. Chua and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," in IEEE Access, vol. 2, pp. 652-687, 2014, doi: 10.1109/ACCESS.2014.2332453.
- Acharjya, Debi Prasanna, and Kauser Ahmed. "A survey on big data analytics: challenges, open research issues and tools." *International Journal of Advanced Computer Science and Applications* 7.2 (2016): 511-518.
- Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science. 2014 Mar 14;343(6176):1203-5. doi: 10.1126/science.1248506. PMID: 24626916.
- Oussous, A., Benjelloun, F., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431-448. https://doi.org/10.1016/j.jksuci.2017.06.001
- Inoubli, Wissem, et al. "An experimental survey on big data frameworks." Future Generation Computer Systems 86 (2018): 546-564.

Module 2: Apache Hadoop Overview

Key concepts

- Hadoop
- Hadoop Distributed File System (HDFS) Overview
- Using Hadoop Distributed File System
- Cloud Object Storage for Big Data
- Yet Another Resource Negotiator (YARN)

Readings

- *Hadoop Web Site:* Apache Hadoop. (n.d.). Retrieved January 6, 2024
- Apache Hadoop 3.4.0-SNAPSHOT – HDFS Architecture, Retrieved January 7, 2024.
- K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 2010, pp. 1-10, doi: 10.1109/MSST.2010.5496972.
- K. Shvachko, "The exabyte club: LinkedIn's journey of scaling the Hadoop Distributed File System, ", 2021, Accessed January 7, 2024
- Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google file system. In Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03). Association for Computing Machinery, New York, NY, USA, 29–43.
- *Apache Hadoop 3.4.0-SNAPSHOT – Overview*, Retrieved January 67 2024.
- Cloud Object Storage - Amazon S3 - AWS, Amazon Web Services, Inc. Retrieved January 6, 2024.
- What is Amazon S3? - Amazon Simple Storage Service, Retrieved January 6, 2024.
- Cloud Data Storage Deep Dive: S3, GCS, and Azure Blob Storage Compared | Airbyte, Retrieved January 6, 2024.
- Wankhede, Pallavi, Minaiy Talati, and Rutuja Chinchamalatpure. "Comparative study of cloud platforms-microsoft, azure, google cloud platform, and amazon EC2." J. Res. Eng. Appl. Sci 5.02 (2020): 60-64.
- Apache Hadoop 3.4.0-SNAPSHOT – Apache Hadoop YARN, Retrieved January 6, 2024.
- Vavilapalli, Vinod Kumar, et al. "Apache hadoop yarn: Yet another resource negotiator." Proceedings of the 4th annual Symposium on Cloud Computing. 2013.

Module 3: Apache Hadoop MapReduce

Key concepts

- The Path to MapReduce
- MapReduce Overview
- Map Reduce Concepts
- MapReduce Examples
- MapReduce Programming
- MapReduce Optimization

Readings

- Sakr, Sherif, Anna Liu, and Ayman G. Fayoumi. "The family of map reduce and large-scale data processing systems." ACM Computing Surveys (CSUR) 46.1 (2013): 1-44.
- Doulkeridis, Christos, and Kjetil Nørvåg. "A survey of large-scale analytical query processing in MapReduce." The VLDB journal 23 (2014): 355-380.

- Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107–113. Bottom of Form
- What is Amazon EMR? - Amazon EMR. (n.d.). Retrieved January 6, 2024.
- Data-Intensive Text Processing with MapReduce Jimmy Lin and Chris Dyer (University of Maryland) Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 7), 2010.
- Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive data sets. Cambridge university press, 2020.
- Apache Hadoop 3.4.0-SNAPSHOT - MapReduce Tutorial, Retrieved January 6, 2024.
- Yelp, GitHub - Yelp/mrjob: Run MapReduce jobs on Hadoop or Amazon Web Services, GitHub, Accessed January 6, 2024.
- mrjob — mrjob v0.7.4 documentation, Retrieved January 6, 2024.

Module 4: Apache Spark (Part 1)

Key concepts

- Spark Overview
- Spark Components
- Concepts
- Creating Spark DataFrames
- Defining Spark SChemas

Readings

- Salloum, S., Dautov, R., Chen, X. et al.Big data analytics on Apache Spark. Int J Data Sci Anal 1, 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9
- Armbrust, Michael, et al. "Spark sql: Relational data processing in spark." Proceedings of the 2015 ACM SIGMOD international conference on management of data. 2015.
- "*Learning Spark*" *pp. 6 – 17, 22-31, 43-54, 58-61*
- DataFrameReader functions here

Module 5: Apache Spark (Part 2)

Key concepts

- Transformation – Rows
- Transformations – Columns
- Transformations – Join
- Transformations – Aggregations
- Transformations – Working with Null Values
- Transformations – Spark SQL
- Transformations – Caching
- Actions
- Actions – Writing Data

Readings

- Built-in (pyspark.sql.functions) Functions

    - col(col)
    - concat(*cols)
    - lit(col)
    - count(col)
    - max(col)
    - min(col)
- DataFrame Function
    - DataFrame.drop(*cols)
    - DataFrame.select(*cols)
    - DataFrame.withColumn(colName, col)
    - DataFrame.withColumnRenamed(existing, new)
    - DataFrame.join(other[, on, how])
    - DataFrame.corr(col1, col2[, method])
    - DataFrame.count()
    - DataFrame.cov(col1, col2)
    - DataFrame.groupBy(*cols)
    - DataFrame.summary(*statistics)
    - DataFrame.dropna([how, thresh, subset])
    - DataFrame.fillna(value[, subset])
    - DataFrame.createOrReplaceTempView(name)
    - DataFrame.cache()
    - DataFrame.unpersist([blocking])
    - DataFrame.collect()
    - DataFrame.count()
    - DataFrame.first()
    - DataFrame.head([n])
    - DataFrame.show([n, truncate, vertical])
    - DataFrame.take(num)
- Column Function
    - alias(*alias, **kwargs)
    - cast(dataType)
    - Column.eqNullSafe(other)
- GroupedData Function
    - GroupedData.agg(*exprs)
    - GroupedData.avg(*cols)
    - GroupedData.count()
    - GroupedData.max(*cols)
    - GroupedData.mean(*cols)
    - GroupedData.min(*cols)
    - GroupedData.sum(*cols)
- SparkSession Function
    - ` SparkSession.sql(sqlQuery[, args])

Module 6: Big Data Streaming and Design Patterns

Key concepts

- Stream Ingestion and Processing (Part 1)
- Stream Ingestion and Processing (Part 2)
- Analytic Cluster Pattern
- Data Lake Pattern
- Lambda Architecture

Readings

- *Apache Kafka*. (n.d.-b). Apache Kafka. Retrieved February 10, 2024.
- Apache Kafka. (n.d.-c). Apache Kafka. Retrieved February 10, 2024.
- Narkhede, N., Shapira, G., & Palino, T. (2017). KAFKA: The Definitive Guide: Real-Time Data and Stream Processing At Scale. O'Reilly : 1st ed., 297 pages.
- Structured Streaming Programming Guide - SPARK 3.5.0 Documentation. (n.d.). Retrieved February 10, 2024.
- R. Hai, C. Koutras, C. Quix and M. Jarke, "Data Lakes: A Survey of Functions and Systems," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 12, pp. 12571-12590, 1 Dec. 2023.
- Ravat, Franck and Zhao, Yan Data Lakes: Trends and Perspectives. (2019) In: International Conference on Database and Expert Systems Applications (DEXA 2019), 26 August 2019 - 29 August 2019 (Linz, Austria).
- M. Kiran, P. Murphy, I. Monga, J. Dugan and S. S. Baveja, "Lambda architecture for cost-effective batch and speed big data processing," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 2785-2792.
- Martinekuan. (n.d.). Big data architectures - Azure Architecture Center. Microsoft Learn. Retrieved February 10, 2024.

Module 7: NoSQL Database

Key concepts

- Using Databases for Big Data Storage
- NoSQL Database Concepts (Part 1)
- NoSQL Database Concepts (Part 2)
- NoSQL Database Classifications (Part 1)
- NoSQL Database Classifications (Part 2)

Readings

- Davoudian, A., Chen, L., & Liu, M. (2018). A survey on NoSQL stores. ACM Computing Surveys (CSUR), 51(2), 1-43.
- Corbellini, A., Mateos, C., Zunino, A., Godoy, D., & Schiaffino, S. (2017). Persisting big-data: The NoSQL landscape. Information Systems, 63, 1-23.
- Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A. & Gruber, R. E. (2006). Bigtable: A Distributed Storage System for Structured Data. OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006 (p./pp. 205--218).
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P. & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. SIGOPS Oper. Syst. Rev.

- James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. 2013. Spanner: Google's Globally Distributed Database. ACM Trans. Comput. Syst. 31, 3, Article 8 (August 2013), 22 pages.
- Benoit Dageville, Thierry Cruanes, Marcin Zukowski, Vadim Antonov, Artin Avanes, Jon Bock, Jonathan Claybaugh, Daniel Engovatov, Martin Hentschel, Jiansheng Huang, Allison W. Lee, Ashish Motivala, Abdul Q. Munir, Steven Pelley, Peter Povinec, Greg Rahn, Spyridon Triantafyllis, and Philipp Unterbrunner. 2016. The Snowflake Elastic Data Warehouse. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16). Association for Computing Machinery, New York, NY, USA, 215–226.
- Distributed hash table. (2024, February 4). In Wikipedia.
- Gupta, A., Tyagi, S., Panwar, N., Sachdeva, S., & Saxena, U. (2017, October). NoSQL databases: Critical analysis and comparison. In 2017 International conference on computing and communication technologies for smart nation (IC3TSN) (pp. 293-299). IEEE.
- Sharma, S. (2015). An extended classification and comparison of nosql big data models. arXiv preprint arXiv:1509.08035.
- *Polyglot Persistence*. March 18, 2024.
- E. Brewer, "CAP twelve years later: How the "rules" have changed," in Computer, vol. 45, no. 2, pp. 23-29, Feb. 2012.
- Kleppmann, M. (2015). A Critique of the CAP Theorem. arXiv preprint arXiv:1509.05393.

Module 8: Key-Value, Wide-Column and Document Stores

Key concepts

- Key-Value NoSQL Database Systems: Redis & Memcache (Dynamo)
- Wide-Column NoSQL Database Systems: HBase & Cassandra (BigTable)
- Document NoSQL Database Systems: MongoDB & CouchDB

Readings

- DeCandia, Giuseppe, et al. "Dynamo: Amazon's highly available key-value store." *ACM SIGOPS operating systems review* 41.6 (2007): 205-220.
- Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." *ACM Transactions on Computer Systems (TOCS)* 26.2 (2008): 1-26.
- Lakshman, Avinash, and Prashant Malik. "Cassandra: a decentralized structured storage system." *ACM SIGOPS operating systems review* 44.2 (2010): 35-40.
- The Battle of the NoSQL Databases – Comparing MongoDB and CouchDB, Mani Yangkatisal, Published July 3, 2020
- DB-Engines - Document Stores, Retrieved September 10, 2024

Course Structure and Learning Activities

There are 8 content modules in this course and each module may take about 9-12 hours to complete. The final module consists of your final exam for the course.

This course is comprised of the following elements:
Learning Activities:

- **Readings**: Each module may include several required and/or supplemental readings.
- **Video Lessons**: In each module, the concepts you need to know will be presented through a collection of short videos. You may stream these videos for playback within the browser by clicking on their titles.
- **Practice Quizzes**: Each module will include some practice quizzes, intended for you to assess your understanding of the topics. You will be allowed unlimited attempts at each practice quiz. There is no time limit on how long you take to complete each attempt at the quiz. These quizzes do not contribute toward your final score in the class.

Assessments:

- **Discussion Forum**: This course has a place for you to interact with other learners about class-related topics.
- **Summative Module Assessments:** Each module will include at least one summative module assessment. You will be allowed one attempt every eight hours for each assessment. There is no time limit on how long you take to complete each attempt at the assessment. Your highest grade will be recorded.
- **Final Assessment:** This course will contain one summative course assessment. Before taking the exam, please make sure you are in a place with reliable internet connection. No retakes will be granted for the lack of internet access. You are in an online program and the use of the Internet is a requirement.

How to Pass This Course

To qualify for a Course Certificate, simply start verifying your coursework at the beginning of the course and pay the fee. Coursera Financial Aid is available to offset the registration cost for learners with demonstrated economic needs. If you have questions about Course Certificates, please see the help topics here.

Also note that this course is part of the Masters of Data Science program offered by Illinois Institute of Technology. By earning a Course Certificate in this course, you are on your way toward earning a Specialization Certificate in this topic. See more information about the program here.

If you choose not to pay the fee, you can still audit the course. You will still be able to view all videos, submit practice quizzes, and view required assessments. Auditing does not include the option to submit required assessments. As such, you will not be able to earn a grade or a Course Certificate.

The following table explains the breakdown for what is required in order to pass the class and qualify for a Course Certificate. You must pass each and every required activity in order to pass this course.

Activity	Required?	Number per Course	Estimated Time per Module	% Required to Pass	% of Total Grade
Lecture Videos	Yes	3-6 per module	.5-1 hour	N/A	N/A

Practice Quizzes	No	3-6 per module	.5 hour	N/A	N/A
Discussions	No	1 per course	1 hour	N/A	N/A
Summative Module Assessments	Yes	1 per module	.5 hour	80%	7.5%/each module (60%)
Summative Course Assessment	Yes	1 -2 per course	1-3 hours	80%	40%

Letter Grades

Letter grades are used for the final grade. Information about IIT grading system can be found in the Graduate Student Handbook.

Letter Grade	Description	Points	Percentage
A	Excellent	4.00	90-100
B	Above Average	3.00	80-89.99
C	Average	2.00	70-79.99
E	Fail	0.00	Under 70%

Getting and Giving Help

- Use the Learner Help Center to find information regarding specific technical problems. For example, technical problems would include error messages, difficulty submitting assignments, or problems with video playback. If you cannot find an answer in the documentation, you can also report your problem to the Coursera staff by clicking on the *Contact Us!* link available on each topic's page within the Learner Help Center.
- Use the flag icon under each item to report errors in lecture video content, assignment questions and answers, assignment grading, text and links on course pages, or the content of other course materials.
- Familiarize yourself with Coursera's policy on Accessibility.

Academic Integrity

Your attentiveness to academic integrity reflects the value you place on your own work and the work of others. In addition to Coursera's Honor Code, we also have high expectations for conduct during course participation.

Discussion Forums: Expectations

Sharing an online course with other avid learners like you gives you a unique opportunity to share, collaborate, and learn from others and their experiences, and helps you reinforce your understanding of the

topics of the course. Interacting in the Discussion Forums is a great way to engage with your online community. We know that it is not possible to read every discussion forum post, so we recommend that you read those that interest you; and reply when you can contribute. The forum is part of your class activities and everybody is expected to act professionally and be civil and respectful of others in your class. Failure to meet these expectations may be considered a break in the Academic Code of Conduct and may result in your removal from the course. Please, check tips and helpful tools to interact in discussion forums in this document.

Academic Code of Conduct

Above all else, learners are expected to ensure that their conduct helps to create an atmosphere conducive to learning and the interchange of knowledge. While it is understood that some of these items are subject to interpretation, learners should nonetheless endeavor to:

- Be respectful of fellow learners.
- Not discriminate against fellow learners in any manner.
- Conduct peer reviews in a timely manner and give useful feedback on what was done well, helpful suggestions for how to improve, and specific comments about why you gave the grade you chose to assist peers in their learning.
- Turn assignments in on time and follow instructions on all assignments including those affecting the use of technology.
- Be truthful in all communication, which includes, but is not limited to, avoiding academic dishonesty.

Illinois Institute of Technology Copyright Statement