Storing and Retrieving Data

Syllabus

Instructors: Arash Nourian and Alex Milowski
School of Information
University of California, Berkeley

Course Summary

Data Science depends on data, and a core competency mandated by this reliance on data is knowing effective and efficient ways to manage, search and compute over that data. This course is focused on how data can be stored, managed and retrieved as needed for use in analysis or operations. The goal of this course is provide students with both theoretical knowledge and practical experience leading to mastery of data management, storage and retrieval with very large-scale data sets.

Course Description

This course prepares students to deal with large-scale collections of data as objects to be stored, searched over, selected, and transformed for use. We examine both the background theory and practical application of information retrieval, database design and database management, data extraction, transformation and loading for data warehouses and other analytical and operational applications. We will examine both the more traditional methods of information retrieval and database management as well as new approaches using massively parallel computation (MapReduce/Hadoop) for applications including web search, very large scale classification, and analytics.

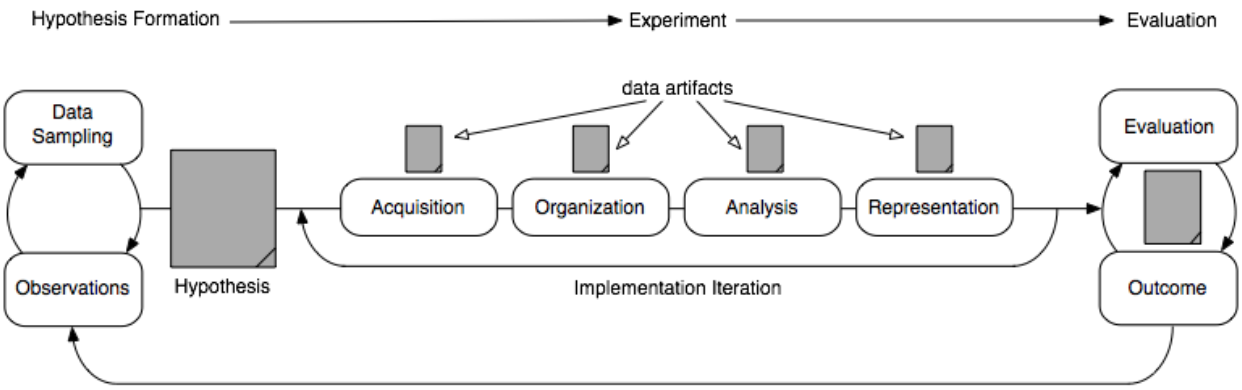
Students will examine through readings, discussion and hands-on experimentation: Theories and methods for data and information organization; database design; searching and retrieval of structured and unstructured information; analysis of relevance, utility; statistical and linguistic methods for automatic indexing and classification; use of parallel processing for information extraction, transformation and loading for search applications; boolean vector and probabilistic approaches to indexing, query formulation, and output ranking; information filtering methods; measures of retrieval effectiveness and retrieval experimentation methodology.

After completing this course students should be able to discuss, plan, and implement storage, search and retrieval systems for large-scale structured and unstructured information systems using a variety of software tools. They should be able to evaluate large-scale information storage and retrieval systems in terms of both efficiency and effectiveness in providing timely, accurate and reliable access to needed information.

Prerequisites

The prerequisites for this course are college-level programming courses in C, Java or Python. Knowledge of database management including SQL is recommended but not required.

Information Organization for Data Science

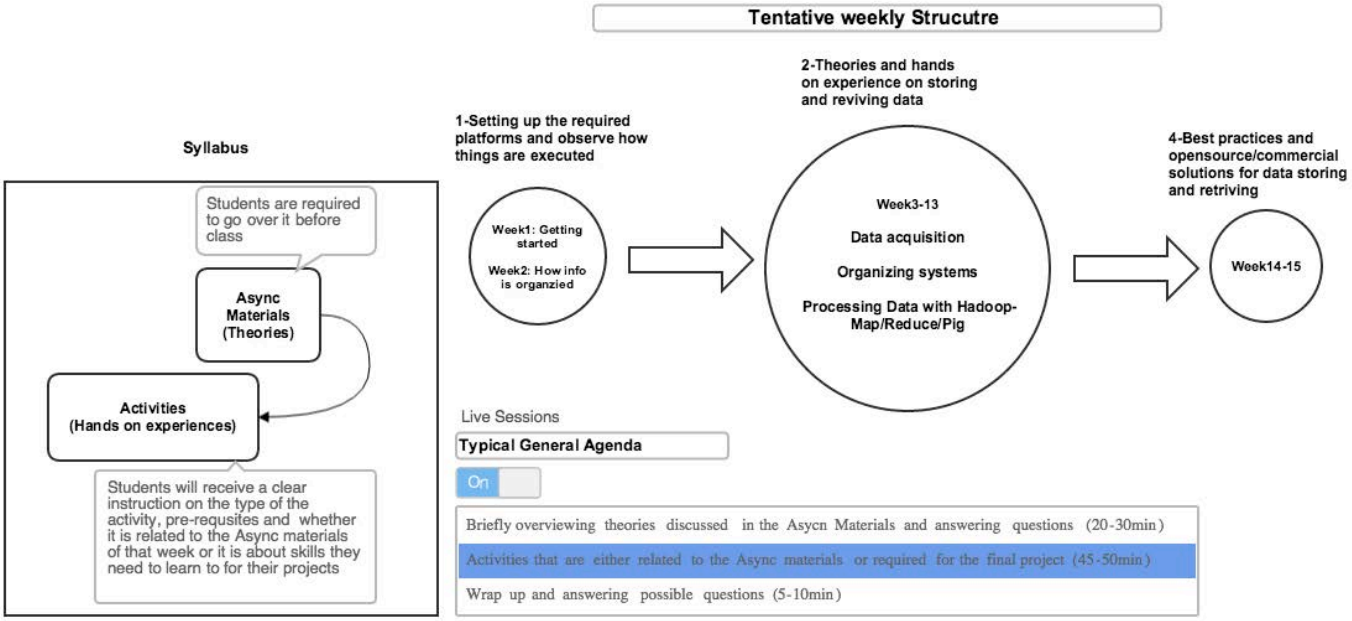


As the course proceeds, you will be expected to do five things on a regular basis:

1. Weekly readings as outlined in the course structure that are expected to be completed pre-session.
2. Weekly pre-session activities that typical start with "scaffolded examples" (e.g. code and instructions) that you should expect to complete **before** the live session. You need to check the Googledoc (URL will be shared) for the activities.
3. Weekly activities **in the live session** that often extend the previous item.
4. Graded assignments relating to theories and live sessions activities.
5. A final team project with a set of milestones presentations given by your team throughout the semester.

Not all of the above items will be expected every single week. Yet, you should endeavor to stay current with the materials and activities. Otherwise, the live-session activities may not produce successful outcomes.

We will use github to turn in our assignments and you will become familiar with the process of branching and pull requests. The pre-session activities will be turned in via github and/or extended through the live session activity. These activities will count towards your class participation.



Grading

20% Class Participation — live session attendance and activities
30% Assignments — graded assignments (see below)
10% Milestone Presentations — final project presentations given in the live-session. *40% Final Project* — the final project and report

Assignments

The details of each assignment are provided separately but a tentative summary is below:

1. Getting Started — An exemplar process designed to walk you through the four stages and get you setup on AWS and github.
2. Acquiring Data — develop a data acquisition program for acquiring data off the Web.
3. Organizing Information — organizing raw acquired information in various storage and database mediums
4. Organizing for Analytics — develop an example text processing application using EMR
5. Organizing for Representation — organizing information from analytics for use in various visualization tools.

Final Project

Final projects teams of 3-4 students are formed and a data science project is selected by the team members. The live sessions and first milestone are used to facilitate team formation. **Milestones**

Milestones are spread throughout the semester and are a short 10 minute/5 slide presentation on the following topics:

1. Project/self pitch — pitch a project or yourself to facilitate team formation.
2. Data acquisition and organization strategy — how will you acquire your data and what information organization challenges do you face?
3. Implementation architecture — what is your implementation strategy and architecture for your solution?
4. Final presentation — the final presentation given to the class at the end of the semester.

Milestones help you keep pace with your project and inform your classmates of the various challenges you face. They will help you collaborate with your cohort and collectively solve various problems you might face.

Project Selection

Projects are selected by the each project team and approved by the instructor. They **should be approved prior to milestone #2** (i.e., *Data acquisition and organization strategy*).

A project must select attempt to process data along some dimension of the *"three V's*" (i.e., Volume, Velocity, or Variety). A good project has more than one aspect of these qualities.

The amount of data processed isn't a hard requirement but there should be sufficient ability to process large amounts of data where the quantification of "large" depend on the complexity of the problem. A project should set out to demonstrate the scalability of the solution rather than to completely process the data acquired. This is simply because a semester may not be enough time to completely acquire and process all the possible data.

Guidance will be provided on project selection so that they produce good outcomes and can be completed within a semester.

To complete a project selection, a **proposal must be sent** to the instructor with sufficient detail of the problem being addressed and the supporting research that data can be acquired. Projects can and do get adjusted as nuances or issues are found that can't be overcome within a semester time period. This is accomplished through communication and mutual agreement with the project team and instructor.

The **proposal is not graded** but should be viewed as the critical step towards the final project selection and good material for inclusion in the final report.

Final Report

A final project report, supporting code, documentation, and possibly a running application is required to be turned in at the end of the semester. The report is intended to be lightweight and 7-20 pages in length covering the following topics:

1. Overview of the problem being addressed.
2. Acquisition and organization of information for analytics.
3. The overall architecture of the solution and necessary implementation details.
4. The results of the project.
5. A retrospective on the project and suggestions for improvements.

Books and Readings:

There are a number of books we'll be using for this course. All of them are available online through the UC Berkeley library proxy or via download.

Required:

The Discipline of Organizing - Glushko Bad
Data Handbook - O'Reilly - McCallum

Optional:

Doing Data Science - O'Reilly - O'Neil & Schutt
Data Science for Business - O'Reilly - Provost & Fawcett
Machine Learning for Hackers - O'Reilly - Conway & White

Course Structure and Readings

This 15 week course consists of follwing sections as part of theories described below. The structure for activities are provided separetly in a shared GoogleDoc file for each week.

I. Introduction to Storing and Retrieving Data

What is Data, how does it differ from information? Knowledge? Why do we care about how and where data is stored? What is the difference between structured data, semistructured, and unstructured data?

- **Week 1: Data, Information and Knowledge and Getting Started with Amazon AWS**
  - Data, Information and Knowledge
  - How Much Data/Information - What is Big?
  - Scales of Information
  - Notions of Relevance, Utility and Applicability
  - Relevance, Utility and Applicability
  - Overview of Information Retrieval
  - Overview of Database Management
  - Overview of MapReduce and Massively Parallel Approaches to Data
  - Data Curation and Applications

- Readings:
  - Ferrucci, D., Brown, E., Chu-Carooll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . Welty, C. (2010, Fall). Building Watson: An overview of the DeepQA project. *AI Magazine*, 59–79.
  - Leskovec, J., Rajaraman, A., & Ullman, J. (2011). *Mining of massive datasets*. New York, NY: Cambridge University Press. **Chapter 1**
  - Manning, C. D., Raghavan, P., & Schuetze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press. **Chapter 1**
  - Saracevic, T. (1975, November–December). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 321–343.
  - Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1975, reprinted 1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management, 26*(6), 766–776.
  - Stonebraker, M. (2009, June 30). The end of a DBMS era (might be upon us). ACM Blogs.
  - Vu, L. (2012, July 11). Getting value from a trillion electron haystack. *iSGTW*.
  - *Journal of the American Society for Information Science, 45*(3) (1994, April). This journal is available via the UCB Library Proxy.

Week 2: Observing the Process of Acquiring and Organizing Information

- Intro: Event-Based Data
- Drinking From Firehoses
- Observing and Experimenting the Process of Acquring Data

- Readings:
  - Chapters 1: Setting the Pace: What is Bad Data? — Bad Data Handbook — Q. Ethan McCallum
  - Chapters 2: Is it Just Me, or Does This Data Smell Funny? — Bad Data Handbook — Q. Ethan McCallum

II. Working with Data and IR Collections

- **Week 3: Big Data Problems**
  - Introduction to Big Data problems
  - The IR problems
  - IR vs Question Answering vs Analytics
  - High Velocity Data
  - High Volume Data
  - High Variety Data - Structured and Unstructured data - Text vs. Binary
  - Making Sense of Big Data
  - Data Structure and Data Maintenance
  - Scale revisited - How big is big data?

- Readings

  - Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.
  - Download **this week's chapter**: Jim Gray on eScience: A Transformed Scientific Method Jagadish, H. V., et al. "Big Data and Its Technical Challenges". *Communications of the ACM v. 57, n. 7 (July 2014)*.

- **Week 4: Data and Text Mining**
  - Intro: Data and Text Mining
  - Data and Text Mining Applications
  - Web-Scale Data Extraction and Retrieval
  - Digital Libraries
  - Web Crawling and Indexing

- Readings:
  - Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003, October). The Google File System. *SOSP'03*.
  - Leskovec, J., Rajaraman, A., & Ullman, J. (2011). *Mining of massive datasets*. New York, NY: Cambridge University Press. **Chapter 4**
  - Lin, J., & Dyer, C. (2010, April). Data-intensive text processing with MapReduce. Manuscript to appear in *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

- **Week 5: Text Processing and Language Properties**
  - Introduction to Week 5
  - Character of Text Collections
  - Term and Language Properties
  - Zipf Distribution
  - Other Distributions
  - Controlled and Uncontrolled Data
  - Controlled and Uncontrolled Data Applications
  - Structured and Semistructured Data Applications

- Readings

  - Daniel Jurafsky & James H Martin. (2nd edition). "*Speech and Language Processing*". **Chapter 2.1**

- **Week 6: Part 1: Data and File Structures**
  - Intro: Data and File Structures
  - Data and File Structures
  - Inverted Files
  - Distributed File Systems
  - Hashed Files
  - RDBMS File Structures
  - NoSQL File Structures
  - Other Specialized File Structures

- **Week 6 Part 2: Database Design**
  - Introduction to DB Design
  - Database Design Process (Revisited)
  - Relational Databases and Document Collections
  - Database Planning and Modeling
  - Data Cleansing and Structuring
  - Relational Normalization
  - DeNormalization and Performance Issues

- Readings
  - Hoffer, J. A., Ramesh, V., & Topi, H. (2012). *Modern database management* (11th ed.). Upper Saddle River, NJ: Prentice Hall (Pearson Educational).
    Leskovec, J., Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. New York, NY: Cambridge University Press. Revisiting Chapter 1
  - Kryder, M. H., & Kim, C. S. (2009, October). After hard drives—What comes next? *IEEE*.
  - Stonebraker, M. (2010, April). SQL databases v. NoSQL databases. *Communications of the ACM, 53*(4), 10–11.

III.Models of Information Storage and Retrieval Systems.

- **Week 7: ISAR System Models**
  - Introduction to ISAR System Models
  - Approaches to Storage and Retrieval
  - Relational DB and the Relational Model
  - Basic IR System Architecture
  - NoSQL Database
  - Flat File and Text Data
  - Data, Metadata and Access Systems - Overhead of Structure
  - Boolean IR operations

- Readings
  - Cattell, R. (2010, December). Scalable SQL and NoSQL data stores. *SIGMOD Record, 39*(4).
  - Manning, C. D., Raghavan, P., & Schuetze, H. (2008). *Introduction to information retrieval*. New York: NY: Cambridge University Press. (Chap. 1)
  - Seltzer, M. (2005, April). Beyond relational databases. *ACM Queue*.

- **Week 8: Information Search Ranking**
  - Introduction to Search Ranking
  - The Vector Space Model
  - Vector Space Ranking
  - Probabilistic Models
  - Probabilistic Ranking
  - Language Models and Ranking
  - Document Ranking - What to Consider
  - Google ranking issues
- Readings
  - Manning, C. D., Raghavan, P., & Schuetze, H. (2008). *Introduction to information retrieval*. New York: NY: Cambridge University Press.
  - Jimmy Lin and Chris Dyer, ***Data-Intensive Text Processing with MapReduce***, Morgan and Claypool, 2010. Free eBook Download this week's **chapter 4**: Inverted Indexing for Text Retrieval

- **Week 9: Big Data Technology**
  - Introduction to Big Data Technology
  - Structured Data Records
  - Row vs Column in DBMS
  - Massively Parallel Processing and MapReduce
  - Hadoop
  - Spark
  - Impala
  - MapReduce vs. Parallel DBMS vs. IR systems
  - Building File Structures with Hadoop and Pig
- Readings
  - Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., . . . Gruber, R. E. (2006). Bigtable: A distributed storage system for structured data. *OSDI 2006*.
  - Dean, J., & Ghemawat, S. (2010, January). MapReduce: A flexible data processing tool. *Communications of the ACM, 53*(1).
  - Ghemawat, S. (2003, October). The Google File System. In *SOSP'03*.
  - Hiemstra, D., & Hauff, C. (2010). MapReduce for information retrieval evaluation: "Let's quickly test this on 12 TB of data." In M. Agosti et al. (Eds.), *CLEF 2010,* LNCS 6360, 64–69.
  - Hiemstra, D., & Hauff, C. (2010). University of Twente at TREC 2010: MapReduce for experimental search. In *TREC19 Proceedings*. Gaithersburg, MD: NIST.
  - Lin, J., & Dyer, C. (2010). *Data-intensive text processing with MapReduce*. Morgan & Claypool.
    Download this week's **chapter 2**: MapReduce Basics
    Download this week's **chapter 7**: Limitations on MapReduce
  - Stonebraker, M., Abadi, D., DeWitt, D. J., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. (2010, January). MapReduce and parallel DBMSs: Friends or foes? *Communications of the ACM, 53*(1).

IV. Automatic Indexing and Information Extraction and Summarization.

- **Week 10: Automatic Indexing**
  - Intro: Automatic Indexing
  - Why indexing? - Indexing goals
  - Indexing Documents vs. Passages
  - Indexing Documents vs. Passages
  - Indexing Documents vs. Passages
  - Stoplists
  - Stemming and Morphological Analysis
  - NLP and Part-of-Speech Tagging
  - Phrase Recognition
  - Disambiguation
  - Using and Inferring Structural Elements
  - Text Location in Indexes
- Readings
  - Salton, G. (1981, Fall). A blueprint for automatic indexing. ACM SIGIR Forum in Cornell University's newsletter, *16*(2), 22–38.

- **Week 11: Information Extraction**
  - Information Extraction
  - Indexing in the Cloud
  - What Does Indexing Entail from Databases, Text, Etc.
  - Using Pig/Hadoop for Index Generation
  - Pig for Indexing
  - Hadoop for Search Indexing
  - Text Mining
  - Indexing and Performance
  - Compressing Indexes
  - Index Sharding
  - Automating Linking
- Readings
  - Olston, C., Reed, B., Srivastava, U., Kuma, R., & Tomkins, A. (2008, June). Pig Latin: A not-so-foreign language for data processing. *SIGMOD'08,* Vancouver, BC, Canada.
  - Pavlo, A., Paulson, E., Rasin, A., Abadi, D. J., DeWitt, D. J., Madden, S., & Stonebraker, M. (2009, June–July). A comparison of approaches to large-scale data analysis. *SIGMOD'09*, Providence, Rhode Island.

- **Week 12: Knowledge Discovery from Data**
  - Intro: Knowledge Discovery from Data
  - Discovering Content From Data
  - Topic Models for Text
  - Text Mining and Inference
  - Classification and Categorization
  - Automatic Classification and Clustering
  - Overview of Machine Learning for Data Mining
  - What Approach for Which Data?
- Readings:
  - Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, Fall). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3).
  - Pazzani, M. J. (2000, March/April). Knowledge discovery from data?. *IEEE Intelligent Systems*.

V. Data Warehouses, Data Marts and ETL Operations

- **Week 13: Data Warehouses, Data Marts, and ETL Operations**
  - Intro: Data Warehouses, Data Marts, and ETL Operations
  - Operational Data vs. Warehousing of Data
  - Mapping Data into a Data Warehouse/ETL for Warehouses and Data Marts
  - Methods for Data Extraction
  - Data Wrangler and Data Wrangling
  - *Data extraction, transformation and loading*
  - Using Pig/Hadoop for ETL
  - Pig for ETL
  - Introduction to practicum - Extracting a data warehouse from Wikipedia
  - Planning and Methods for Transforming Semistructured Data to Structured Data
  - Data Normalization and Validation

- Readings:
  - Inmon, W. H. (2000). Building the data warehouse: Getting started.
  - Meijer, E., & Bierman, G. (2011, March). A co-relational model of data for large shared data banks. *Programming Languages, 9*(3).

VI. Event Based Data

- **Week 14: Event-Based Data**
  - The Twitter Stream
  - Processing Twitter with Spark
  - Approaches to Dealing with Continuous Change
  - Streambase Systems and Events
  - Volt DB
  - Processing Log Data
  - The 3Vs Revisited
  - Really big data - MReplay and Ark.com

- Readings:
  - Abadi, D. J. et al. (2003). Aurora: A new model and architecture for data stream management. *The VLDB Journal, 12*(2), 120–139.
  - Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., Seidman, G., . . . Zdonik, S. (2002). Monitoring streams—A new class of data management applications. *Proceedings of the 28th VLDB Conference* Hong Kong.
  - Lin, J. (2013, March). MapReduce is good enough? *Big Data, 1*(1). Mary Ann Liebert, Inc.

VII. Summary and Wrap-up

- **Week 15: Final Reports**
  - Putting it All Together
  - Which Technology for Which Problem?
  - How this Course Will Impact Your Other Courses?
  - Crystal Ball Gazing - Future Trends in Data Storage and Retrieval