MIDS W205

Storing and Retrieving Data

Instructors:

Jari Koister, jari.koister@ischool.berkeley.edu

Dan McClary, dan.mcclary@ischool.berkely.edu

Karthik Ramasamy, karthik.ramasamy@ischool.berkeley.edu

Course Overview

Storing, managing, and processing datasets is foundational to both applied computer science and data science. Indeed, successful deployment of data science in any organization is closely tied to how data is stored and processed. This course introduces the fundamentals of data storage, retrieval and processing systems. As these fundamentals are introduced, exemplary technologies will be used to illustrate how storage and processing architectures can be constructed.

This course aims to provide a set of "building blocks" by which one can construct a complete architecture for storing and processing data. The course will examine how technical architectures vary depending on the problem to be solved, the reliability and freshness of result. The problems are being considered in the context of data analytics. The course considers traditional architectures as well as so called big data architectures. Students should consider both small and large data sets as both are equally important both justifying different tradeoffs. Exercises and examples will consider both simple and complex data structures, as well as data ranges from clean and structured to dirty and unstructured.

Prerequisites

- Previous experience with Python
- An understanding of algorithmic complexity (e.g. "Big O" notation)

Evaluation

1. 12 weekly labs (weeks 1-12): 15% of grade
2. 2 Exercises, spanning weeks 1-7, and 8-14: 20% each (40% total)
3. 2 Exams, mid-term (week 7) and final: 45%

Required and Recommended Reading

Week 1

Required reading:

[1] Jeff Hammerbacher, Information Platforms and the Rise of the Data Scientist. Beautiful Data: The Stories Behind Elegant Data Solutions.O'Reilly, Chapter 5. July 2009. link

[2] Jari Koister. Dimensions for Characterizing Analytics Data Processing Solutions. White Paper for DATASCI W205, 2015. link

[3] Jianwei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. Chapter 1, Pages 1-35. Third Edition, Morgan Kaufman, 2012.

Recommended (but not required) reading:

[4] DJ Patil and Hilary Mason. Data Driven: Creating a Data Culture. link

Week 2

Required reading:

[5] Sriram Krishna and Eva Tse. Hadoop Platform as a Service in the Cloud. Netflix blog post 2013. link

[6] Nathan Marz and James Warren. Big Data: Principles and best practices of scalable real-time data systems. Sections 1.4 - 1.10, Manning 2015.

Week 3

Required reading:

[7] H.A. Proper. Data Schema Design as a Schema Evolution Process. Data & Knowledge Engineering, 22(2):159-189, 1997.

[8] E.F. Codd. A Relational Model of Data for Large Shared Data Banks. ACM Information Retrieval. 13(6): 377-387, 1970

[9] P. Chen. The Entity-Relationship Model -- Toward a Unified View of Data. ACM Transactions on Database Systems. 1(1): 9-36, 1976.

Week 4

Required reading:

[10] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. SOSP'03, October 19–22, 2003, Bolton Landing, New York, USA. link

[11] Jay Kreps. The Log: What every software engineer should know about real-time data's unifying abstraction, LinkedIn Blog 2013. link

Week 5

Required reading:

[12] Panos Vassiliadis, A Survey of Extract–Transform– Load Technology. International Journal of Data Warehousing & Mining, 5(3), 1-27, July-September 2009. link

[13] Bryce Allen, John Bresnahan, Lisa Childers, Ian Foster, Gopi Kandaswamy, Raj Kettimuthu, Jack Kordas, Mike Link, Stuart Martin, Karl Pickett, and Steven Tuecke. Software as a Service for Data Scientists. february 2012 | vol. 55 | no. 2 | communications of the acm. link

Recommended (but not required) reading:

[14] Jay Kreps et al. Kafka: a Distributed Messaging System for Log Processing.NetDB'11, Jun. 12, 2011, Athens, Greece.ACM 978-1-4503-0652-2/11/06. link

Week 6

Required reading:

[15] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker and I. Stoica. Spark: Cluster Computing with Working Sets, HotCloud 2010, June 2010.

[16] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008

Week 7

Required reading:

[17] Graefe, Goetz. "Query evaluation techniques for large databases." ACM Computing Surveys (CSUR) 25.2 (1993): 73-169.

[18] Chaudhuri, Surajit. "An overview of query optimization in relational systems." *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. ACM, 1998.

Recommended (but not required) reading:

[19] Stonebraker, Mike, et al. "C-store: a column-oriented DBMS." Proceedings of the 31st international conference on Very large data bases. VLDB Endowment, 2005. *[OPTIONAL]*

Week 8

Required reading:

[20] S. S. Stevens. On Theory of Scales and Measurement.Science. Vol. 103, No. 2684. 1946. link

[21] John W. Tukey. We need both exploratory and confirmatory. The American Statistician, Vol. 34, No. 1 (Feb., 1980), pp. 23-25. link

[22] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis Dremel: Interactive Analysis of Web-Scale Datasets. Proceedings of the VLDB Endowment, Vol. 3, No. 1. 2010. link

Week 9

Required reading:

[23] Ankit et.al. Storm@Twitter, Proceedings of SIGMOD Conference, 2014. link

[24] Sanjeev et.al. Twitter Heron: Streaming at Scale, Proceedings of SIGMOD Conference, 2015. link

Week 10

Required reading:

[25] Elmagarmid, A., Ipeirotis, P., & Verykios, V. (2007). Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 19, 1, 1-16 link

[26] Erhard Rahm, Hong Hai Do. Data Cleaning: Problems and Current Approaches, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, IEEE 2000. link

[27] Wang, Richard Y., and Diane M. Strong. "Beyond accuracy: What data quality means to data consumers." *Journal of management information systems*(1996): 5-33. link

Recommended (but not required) reading:

[28] Jianwei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. Chapter 3, Pages 83-120. Third Edition, Morgan Kaufman, 2012.

Week 11

[29] Amaral, LAN, Scala, A, Barthelemy, M, Stanley, HE. "Classes of Small World Networks." *Proc. Natl. Acad. Sci. U. S. A.* 97, 11149-11152 (2000). link

[30] Liljeros, F, Edling, CR, Amaral, LAN, Stanley, HE, Aberg, Y. "The web of human sexual contacts." *Nature* 411, 907-908 (2001). link

[31] M. E. J. Newman, "The structure of scientific collaboration networks." *Proc. Natl. Acad. Sci. USA* **98**, 404-409 (2001). link

Weeks 12-14

Required reading:

[32] Jianwei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. Chapter 5, Pages 83-120. Third Edition, Morgan Kaufman, 2012.

Other selected readings may be assigned.

Course Topics

Week 1 : Course Introduction.

We will introduce data driven organizations and why the needs for storing and retrieving data is changing. We will also introduce a simple model for characterizing data and processing needs.

- Introduction: Data Driven Organizations
- Concepts: Dimension for Data
- Concepts: Dimensions for Processing

Week 2 : Dimensions and Scaling: Understanding Tradeoffs

In this module we will provide an intuition for data size, storage access performance and processing needs. We will also discuss considerations for data transfer. We discuss fundamental architectural concepts such as scale-out, scale-up, and single node versus distributed systems.

- Introduction: Data Size, Transfer
- Introduction: Processing
- Concepts: Data Scaling
- Concepts: Scaling Processing
- Architecture: Single Node vs. Distributed Storage
- Architecture: Single Node vs. Distributed Processing

Week 3 : Structure and Organization

In this module we will describe how data is structure and defined. We will introduce the concepts of schemas and how they are modeled.

- Structuring Data
- Schema
- When is Schema Applied
- Schema as Contract
- Semantic Modeling
- Physical Schema

Week 4 : Data Lakes: Storage and Maintenance

In this module we will introduce how large sets of unstructured data (sometimes called data lakes) can be stored and processed. We discuss different underlying storage solutions, their characteristics and technical underpinnings. We also introduce the concepts of provenance and governance of data.

- What is a Data Lake?
- High-level data systems architectures
- Data Characteristics
- Mapping to Data Architectures.
- Mapping to Data Architectures: NOSQL and HDFS
- Mapping to Data Architectures: Relational/Columnar
- Mapping to Data Architectures: Software Defined Object Storage (Swift, S3)
- Management, Provenance, Governance

Week 5 : Data Ingestion: Storage and Maintenance

Data most come from somewhere. In this module we discuss various solutions for data ingestion and data movement.

- Intro Data Ingestion/Loading
- Traditional ETL/ELT
- Ingestion of high velocity Logs
- Big Data Ingest: Logs + ETL
- Moving large data sets

Week 6 : Data Processing and Aggregation

Data processing such as aggregation, grouping and filtering are fundamental to analytics. In this module we discuss methods for such processing.

- Introduction: Processing
- Methods of Processing
- Functional Programming and Parallelism
- Processing in Stages
- Understanding Aggregation

Week 7 : Querying Data

Queries are the fundamental way of extracting knowledge from of data. In this module we discuss the fundamental principles and methods for querying data.

- Review: Schema, RDBMS, and DAGs
- Motivation for Declarative Languages
- Structured Query Language (SQL)

- Joins
- Analytical SQL and Windows
- Indexes
- View/Partitions
- Approximate SQL

Week 8 : Exploring Data

When we do not know what we are looking for, or what to ask of our data we need to explore it. In this section we present the fundamentals of data exploration and also how to prepare the data for exploration.

- Understanding your data
- Exploratory Data Analysis
- Visualization and its realization
- Example Tool: BDD
- Confirmatory Data Analysis
- Example Tool: BigQuery
- Sampling, Enriching, Merging
- Clustering, Classification

Week 9 : Streaming Data

Analytics on data in motion is becoming increasingly important. In this module we will describe how to build streaming analytics applications using Storm as an example. Storm is one of the leading streaming analytics platforms.

- Storm Overview
- Storm Example
- Storm Architecture
- Storm Deployment
- Storm Issues
- Twitter Heron
- Twitter Heron Performance/Operational Experiences

Week 10 : Cleaning data

Data quality and wrangling is key in any analytics systems. These processes can be very processing intensive. In this module we describe the basic techniques of cleaning data. With this as a base we discuss why these are processing heavy and what can be done about it.

- Defining data quality
- Single Stream Issues
- Missing Values
- Entity linkage
- Record Linkage
- Scaling Record Linkage
- Ontologies and semantics

Week 11 : Graph Models and Analysis

There are many interesting applications of graph based processing model. In this module we will describe bow these work and what the computational implications are for graph processing frameworks.

- Introduction: Defining Graphs
- Degree, Diameter and Components
- Path Finding
- Ranking and Centralities
- Communities
- Storing Graphs

Week 12 : Serving Data

Once data is processed and we have analytics results we want to use them for some purpose. It is important to understand the difference in requirements between analytics processes and how to make data available for users or applications. In this module we will present fundamental ways and considerations for serving data.

- Reporting
- In application analytics
- Serving at scale

Week 13 : Advanced Topics

In this module we will cover a few advanced but very interesting topics. We will provide more depth to stream processing. We will describe processing in preparation for machine learning algorithms. We will also discuss some important considerations with respect to mining data streams. Finally we will introduce the concepts of data cubes, a fundamental technology in data processing and storage.

- Advanced Streaming
- ML Pipelines
- Mining Streams
- Cuboids

Week 14 : Course Wrap Up

In this module we will review what we learnt earlier in the course reiterating some key points. We will also provide interesting interviews with leading data analytics minds.

- Reviews
- Interviews