MIDS W205

Storing and Retrieving Data

Instructors:

Jari Koister, jari@ischool.berkeley.edu

Arash Nourian , nourian@ischool.berkeley.edu

Amit Bhattacharyya , amitbl@ischool.berkeley.edu

Uri Schoenfeld, shuri@ischool.berkeley.edu

Edward Fine, efine@ischool.berkeley.edu

Course Overview

Storing, managing, and processing datasets are foundational to both applied computer science and data science. Indeed, successful deployment of data science in any organization is closely tied to how data are 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 and the reliability and freshness of the 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 datasets because both are equally important, both justifying different trade-offs. Exercises and examples will consider both simple and complex data structures, as well as data ranges from clean and structured to dirty and unstructured.

Course Learning Objectives

- Understand all main architectural components involved in building analytics processes and applications that results from data science activities. This includes for example data definition and storage, data ingestion, data processing querying, data cleaning, data serving.
- Understand fundamental architectural concepts and characteristics that are considered when building analytics processes and applications. This includes data scale, processing complexity, network performance.
- Understand the nature and needs of processing and storage for the various processes involved in data analytics. Be able to evaluate any solution according to some fundamental concepts (dimensions)
- Understand trade-offs between different technology choices. Conceptual model differences, functional differences, scale and performance differences.
- Hands-on experience and introductory knowledge of selected technologies.
- Ability to analyze a problem and select an appropriate architecture based on functional and non-functional requirements as well as known characteristics of technical solutions.

Prerequisites

- Previous experience with Python.
- Basic knowledge of Unix/Linux commands and tools as well as concepts such as processes, file systems.
- An understanding of algorithmic complexity (e.g., "big O" notation)

Evaluation

1. 10 labs (spread through out the course): [25% of grade]
2. 2 exercises, spanning Weeks 1–7 and 8–14: (20% ex 1, 20% ex 2 respectively) [total 40% of grade]
3. Final project: [total 35% of grade]

Required and Recommended Reading

For papers and articles the list below has links for download. If you have trouble downloading them using these links, you can download them from here: https://github.com/jarikoi/interesting-papers

Week 1

Required reading:

[1] Hammerbacher, J. (2009). *Information platforms and the rise of the data scientist. In Beautiful data: The stories behind elegant data solutions*. O'Reilly. link

[2] Koister, J. (2015). Dimensions for characterizing analytics data processing solutions. White paper for DATASCI W205. link

[3] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufman, Chapter 1, pp. 1–35. link

Recommended (but not required) reading:

[4] Patil, D.J., & Mason, H. (2015). *Data driven: Creating a data culture*. link

Week 2

Required reading:

[5] Krishna, S., & Tse, E. (2013). Hadoop platform as a service in the cloud. Netflix blog post. link

[6] Marz, N. & Warren, J. (2015). Big data: Principles and best practices of scalable real-time data systems. Manning, Sections 1.4—1.10.

Week 3

Required reading:

[7] Codd, E. F. (1970). A relational model of data for large shared data banks. *ACM Information Retrieval, 13*(6): 377–387. Link

[8] Chen, P. (1976). The entity relationship model—toward a unified view of SATA. *ACM Transactions on Database Systems, 1*(1): 9–36. Link

Recommended (but not required) reading:

[9] Proper, H. A. (1997). Data schema design as a schema evolution process. *Data & Knowledge Engineering, 22*(2):159–189. link

Week 4

Required reading:

[10] Ghemawat, S., Gobioff, H., & Leung, S. (2003). The Google file system. SOSP'03, October 19–22, Bolton Landing, New York, USA. link

[11] Kreps, J. (2013). The log: What every software engineer should know about real-time data's unifying abstraction. LinkedIn blog. link

Week 5

Required reading:

[12] Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing & Mining, 5*(3), 1–27. link

Recommended (but not required) reading:

[13] Kreps, J. et al. (2011). Kafka: A distributed messaging system for log processing. NetDB'11, Athens, Greece. ACM 978-1-4503-0652-2/11/06. link

Week 6

Required reading:

[14] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *HotCloud*.

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

Week 7

Required reading:

[16] Graefe, G. (1993). Query evaluation techniques for large databases. *ACM Computing Surveys (CSUR)*, *25*(2): 73–169. Link

[17] Chaudhuri, S. (1998). An overview of query optimization in relational systems. Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Link

Recommended (but not required) reading:

[18] Stonebraker, M. et al. (2005). C-store: A column-oriented DBMS. Proceedings of the 31st International Conference on Very Large Databases. VLDB Endowment. *Link*

Week 8

Required reading:

[19] Stevens, S. S. (1946). On theory of scales and measurement. *Science*, *103*(2684). link

[20] Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, *34*(1): 23–25. link

[21] Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., & Dremel, T. V. (2010). Interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, *3*(1). link

Week 9

Required reading:

[22] Toshniwal, A. et al. (2014). Storm@Twitter. Proceedings of SIGMOD Conference. link

[23] Kulkarni, S. et al. (2015). Twitter Heron: Streaming at scale. Proceedings of SIGMOD Conference. link

Week 10

Required reading:

[24] Elmagarmid, A., Ipeirotis, P., & Verykios, V. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1): 1–16. link Read the following sections: 1,2 3 {3.1.1,3.1.2,3.1.4,3.3.1} 4 {4.1,4.3,4.5,4.6,4.8},5{5.1,5.2},7. The rest is optional.

[25] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. link Read the following sections: 1,2 3 {3.1.1,3.1.2,3.1.4,3.3.1} 4 {4.1,4.3,4.5,4.6,4.8},5{5.1,5.2},7. The rest is optional.

[26] Wang, R. Y., &. Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4): 5–33. link Read the following sections: 1,2 3 5. The rest is optional.

Read the following sections: Introduction, Preliminary Conceptual Framework ,Toward a Hierarchical Framework of Data Quality. The rest is optional.

Recommended (but not required) reading:

[27] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*, (3rd ed.). Morgan Kaufman, Chapter 3, pp. 83–120.Read the following sections: 3.1, 3.2, 3.3.1, 3.4.8-3.4.9 Optional : 3.3.2-3.4.7, 3.5

Week 11

[28] Amaral, L. A. N., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Procedures of the National Academy of Science*, 97, 11149–11152. link

[29] Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., & Aberg, Y. (2001). The web of human sexual contacts. *Nature, 411*: 907–908. link

[30] Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science, 98:* 404–409. link

Weeks 12–14

Recommended (but not required) reading:

[31] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*, (3rd ed.). Morgan Kaufman, Chapter 5, pp. 187–194, 210-218.Read the following sections: 5.1 Optional : 5.2- 5.5

[32] Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K., & Tuecke, S. (2012). Software as a service for data scientists. *Communications of the ACM*, 55(2). Link

Other selected readings may be assigned.

Course Topics

Week 1: Course Introduction and Architecture

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

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

Week 2: Dimensions and Scaling: Understanding Trade-Offs

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 are 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 their 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.

- Introduction to Data Ingestion/Loading
- Traditional ETL/ELT
- Ingestion of High-Velocity Logs
- Big Data Ingest: Logs + ETL
- Moving Large Datasets

Week 6: Data Processing and Aggregation

Data processing, such as aggregation, grouping, and filtering, is 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 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

Comment [PH1]: The CMS doesn't have a title for this unit.

Data quality and wrangling are 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 models. In this module we will describe how 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

Comment [PH2]: In the CMS, the title of this unit is "Advanced Topics"

Once data are 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

Comment [PH3]: In the CMS, this unit is titled "Serving Data"

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 learned earlier in the course, reiterating some key points. We will also provide interesting interviews with leading data analytics minds.

- Reviews
- Interviews