◆ Storing and Retrieving Data ◆

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 is stored and processed. This course introduces the fundamentals of data storage, retrieval, and processing systems in the context of common data analytics processing needs. As these fundamentals are introduced, representative technologies will be used to illustrate how to construct storage and processing architectures. 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 course considers the complete breadth of technology choices. The content spans from traditional databases and business warehouse architectures, so-called big-data architectures, to streaming analytics solutions and graph processing. Students will consider both small and large datasets because both are equally important and both justify different trade-offs. Exercises and examples will consider both simple and complex data structures, as well as data that is both clean and structured and dirty and unstructured.

Skills Developed

Analytics Solution Architectures | Data at Scale Concerns and Tradeoffs | Distributed Data Processing | Relational Databases | Graph Databases | Streaming Data Applications | Cube Technology

Grading

- Weekly Labs – 25%
- Weekly Exercises – 40%
- Final Project – 35%

Learning Objectives

Unit 1: Introduction, Scaling

- Discuss why the needs for storing and retrieving data are changing
- Consider dimension and scaling and the relationship between data size, storage needs, and processing needs
- Learn the basic metrics for platform scale and performance

Unit 2: Data Ingestion, Processing, Querying and Exploration

- Outline how data is structured and defined and how schemas are modeled
- Explore emerging analytics architecture for small and big data
- Move and ingest data
- Compare and contrast methods of data processing including aggregation, grouping, and filtering
- Understand fundamental principles and methods for querying data
- Study processing for data exploration

Unit 3: Graph Data, Streaming Data, and Dirty Data

- Build streaming analytics applications
- Cover graph-based processing models
- Run data cleaning processes

Unit 4: Serving Data and Advanced Topics

- Investigate the difference between analytics processes and making data available for users or applications
- Describe processing needs for ML pipelines
- Outline the benefits and limitations of Cube technologies

- Examine sampling and filtering for data streams

Unit 5: Course Reviews, Reflection and Interviews

- Review core concepts
- Delve into interviews with three thought leaders in data analytics