**** **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 are 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, 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 course considers the complete breadth of technology choices. The content spans from traditional databases and business warehouse architectures, so-called big-data architectures, streaming analytics solutions, and graph processing. 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.

**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**

* Understand why the needs for storing and retrieving data are changing
* Considerations of dimension and scaling and the relationship between data size, storage needs, and processing needs
* Understand the basic metrics for platform scale and performance.

**Unit 2: Data Ingestion, Processing, Querying and Exploration**

* Outline how data are structured and defined and how schemas are modeled
* Understand emerging analytics architecture for small and big data.
* Know how to ingesting and move data.
* Compare and contrast methods of data processing including aggregation, grouping, and filtering
* Understand fundamental principles and methods for querying data
* Understand processing for data exploration.

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

* Learn to build streaming analytics applications
* Understand graph-based processing models
* Learn to building data cleaning processes.

**Unit 4: Serving Data and Advanced Topics**

* Understand the difference between analytics processes and making data available for users or applications
* Processing needs for ML pipelines.
* Understanding the benefits and limitations of Cube technologies.
* Sampling and filtering for data streams.

**Unit 5: Course Reviews, Reflection and Interviews**

* Review of core concepts.
* Interviews with 3 thought leaders in Data Analytics.