DS4300: Large-Scale Information Storage & Retrieval

Introduces data and information storage approaches for structured and unstructured data. Shows how to build large-scale information storage structures using distributed storage facilities. Explores data quality assurance, storage reliability, and challenges of working with very large data volumes. Requires use of non-relational, document, key-column, key-value, and graph databases. Teaches how to model multi-dimensional data. Implements distributed databases. Considers multi-tier storage design, storage area networks, and distributed data stores. Applies algorithms, including graph traversal, hashing, and sorting to complex data storage systems. Considers complexity theory and hardness of large-scale data storage and retrieval. Requires programming in R, Python, and C++.

Pre-Requisites: DS3200; DS4100

NU Core Designations: AD

After completing this course, students will be able to:

• classify data storage approaches based on data object type, data retrieval and analysis requirements
• select an appropriate data storage structure depending on object type and analysis goals
• plan an information repository for data analysis, data visualization, and discovery
• implement large-scale non-relational data repositories
• distinguish between storage needs for statistical and non-statistical analysis
• outline tiered information architectures for efficient data retrieval and search
• store social graphs, documents, geographical, non-textual, and time series data
• evaluate distributed data storage and retrieval approaches
• describe hardware and network requirements for large-scale data storage

Achievement of learning outcomes will be assessed through:

• completion of programming assignments in R, Python, and C++
• development of a substantial term project requiring the design and implementation of a storage architecture
• mid-term and final exams