CSE 511: Data Processing at Scale

About this course

Database systems are used to provide convenient access to disk-resident data through efficient query processing, indexing structures, concurrency control, and recovery. This course delves into new frameworks for processing and generating large-scale datasets with parallel and distributed algorithms, covering the design, deployment and use of state-of-the-art data processing systems, which provide scalable access to data.

Specific topics covered include:

- Efficient query processing
- Indexing structures
- Distributed database design
- Parallel query execution
- Concurrency control in distributed parallel database systems
- Data management in cloud computing environments
- Data management in Map/Reduce-based
- NoSQL database systems

Required prior knowledge and skills

- Basic statistics and computer science knowledge including computer organization and architecture, discrete mathematics, data structures, and algorithms
- Knowledge of high-level programming languages (e.g., C++, Java) and scripting language (e.g., Python)

Learning Outcomes

Learners completing this course will be able to:

- Differentiate among major data models such as relational, spatial, and NoSQL
- Perform queries (e.g., SQL) and analytics tasks in state-of-the-art database systems
- Apply leading-edge techniques to design/tune distributed and parallel database systems
- Utilize existing NoSQL database systems as appropriate for specified cases
- Perform database operations (e.g., selection, projection, join, and groupby) in state-of-the-art cluster computing systems such as Hadoop/Spark
- Perform scalable data processing operations (e.g., selection, projection, join, and groupby) in cloud computing environments, including Amazon AWS

Estimated Workload/Time Commitment Per Week

15 - 20 hours per week

Technology Requirements

Hardware - Standard hardware with major OSSoftware and Other (programs, platforms, services, etc.) - To complete course projects, some of the following may be required: Amazon AWS, Cloud, Hadoop/Spark, GitHub, PostgreSQL, MongoDB, Neo4j.