Scalable Data Processing (CSE 511)

Note: Below outline is subject to modifications and updates.

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

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

Projects:
Project 1: Movie Recommendation Database
Project 2: Distributed Movie Recommendation Database
Project 3: Location-Aware Twitter Analytics
Project 4: Spatial Data Processing using Apache Spark
Project 5: SQL queries on Amazon EC2

Course Content

Scalable Data Processing
Lead: Mohamed Sarwat, Ph.D.
Updated 12/28/2017
Instruction
Video Lectures
Other Videos
Readings
Interactive Learning Objects
Live office hours
Webinars

Assessments
Practice activities and quizzes (auto-graded)
Practice assignments (instructor- or peer-reviewed)
Team and/or individual project(s) (instructor-graded)
Final exam (graded)

Estimated Workload/ Time Commitment Per Week
Approximately 9 hours per week

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

Technology Requirements

Hardware
Standard with major OS

Software and Other
To complete course projects, some of the following software may be required: Amazon AWS Cloud, Hadoop/Spark, GitHub, PostgreSQL, MongoDB, Neo4j.

Course Outline

Unit 1: Basic Data Processing Concepts

Learning Objectives
1.1: Explain Data Models and Data processing concepts
1.2: Utilize Relational Model and Relational Algebra
1.3: Utilize SQL query language

- Unit Introduction
- Module 1: Big Data and Data Processing
  o Introduction to Data and Data Processing

Scalable Data Processing
Lead: Mohamed Sarwat, Ph.D.
Updated 12/28/2017
Module 2: Basic Data Concepts

- Database Systems - What and Why?
  - Database Management Systems
  - Data Model
- Database Design: Entity Relationship Model to Relational Model
  - Entity Relational Model
  - ER to Relational Model
  - Assignment: Create a Movie Database
- Relational Model and Relational Algebra
  - Relational Data Model
  - Relational Algebra: Query Language
  - Query Language: Union
  - Query Language: Difference
  - Query Language: Cartesian Product
  - Query Language: Selection
  - Query Language: Projection
  - Query Language: Intersection
  - Query Language: θ-Join
- SQL Query Language:
  - Part 1: SQL Query Language
  - Part 2: SQL Query Language
  - Assignment: SQL Query for Movie Recommendation

Unit 2: Data Storage and Indexing

**Learning Objectives**

2.1 Recognize major data storage layouts
2.2 Identify major indexing schemes in Database Systems

- Unit Introduction
- Module 1: Major Storage Layouts
  - Introduction to Data Storage
  - Alternative File Organizations
- Module 2: Major Indexing Schemes in Database Systems
  - Hash-based Indexes
  - Index Classification
Unit 3: Transactions and Recovery

Learning Objectives
3.1 Examine the ACID properties
3.2 Explain Transactions and Concurrency Control concepts
3.3 Describe how recovery from failures happens in database systems

- Unit Introduction
- Module 1: ACID Properties
  - Principles of Transactions: ACID Properties
- Module 2: Concurrency Control Concepts
  - Concurrency Control
- Module 3: Lock-based Concurrency Control and Recovery from Failures
  - Lock-Based Concurrency Control
  - Database Recovery

Unit 4: Principles of Distributed and Parallel Database Systems

Learning Objectives
4.1 Describe data fragmentation and replication models
4.2 Describe the components of a distributed database
4.3. Apply skills learned to complete an assignment using data partitioning

- Unit Introduction
- Module 1: Distributed Databases: Why, What?
  - Why Distribution?
- Module 2: Data Fragmentation and Replication Model
  - Introduction to Fragmentation
  - Introduction to Replication
  - Assignment: Data Fragmentation
- Module 3: Advanced Distributed Database Systems
  - Query Processing and Optimization in Distributed Databases
  - Distributed Query Processing
  - Total Cost of Query Execution Plan
  - Assignment: Query Processing
- Module 4: Parallel Database Systems
  - Parallel Data Architecture
    - Introduction to Parallel DBMS
Unit 5: NoSQL Database Systems

Learning Objectives

- Unit Introduction
- Module 1: NoSQL Database Systems
  - Key-Value Stores
  - Graph Databases
  - Document Databases
- Module 2: Big Data Analytics Systems
  - Intro Map-Reduce / Spark
  - Data Analytics in Map-Reduce / Spark
  - Graph Processing Engines
- Module 3: Data Processing on Modern HW
  PROJECT: Distributed Movie Recommendation Database

Unit 6: Big Data Tools

PROJECT: Location-Aware Twitter Analytics
PROJECT: Spatial Data Processing using Apache Spark

Unit 7: Data Management in the Cloud

Learning Objectives

7.1 Explain data processing in the cloud
7.2 Evaluate service models
7.3 Evaluate deployment models

- Unit Introduction
- Module 1: Introduction to Cloud Computing
  - Introduction to Cloud Computing
- Module 2: Service Models
Unit 8: Cloud-based Data Management

**Learning Objectives**

8.1 Explain AWS

- Unit Introduction
- Module 1: Amazon Web Services
  - Introduction to Amazon Web Services
  - AWS Computing
  - AWS Storage
  - AWS Queueing Services
- Module 2: Build an Elastic Cloud Application
  - AWS Interfaces
  - Auto-Scaling
- Module 3: Build a MapReduce Cloud Application
  - Scalable Data Processing
  - AWS Security

**PROJECT:** SQL queries on Amazon EC2

**Creators**

Established in Tempe in 1885, Arizona State University (ASU) has developed a new model for the American Research University, creating an institution that is committed to access, excellence and impact.

As the prototype for a New American University, ASU pursues research that contributes to the public good, and ASU assumes major responsibility for the economic, social and cultural vitality of the communities that surround it. Recognizing the university’s groundbreaking initiatives, partnerships, programs and research, *U.S. News and World Report* has named ASU as the most innovative university all three years it has had the category.

The innovation ranking is due at least in part to a more than 80 percent improvement in ASU’s graduation rate in the past 15 years, the fact that ASU is the fastest-growing research university...
in the country and the emphasis on inclusion and student success that has led to more than 50 percent of the school’s in-state freshman coming from minority backgrounds.

Mohamed Sarwat

Mohamed Sarwat is an Assistant Professor of Computer Science and the director of the Data Systems (DataSys) lab at Arizona State University (ASU). He is also an affiliate member of the Center for Assured and Scalable Data Engineering (CASCADE). Before joining ASU, Mohamed obtained his MSc and PhD degrees in computer science from the University of Minnesota. His research interest lies in the broad area of data management systems.

Ming Zhao

Ming Zhao is an associate professor of the ASU School of Computing, Informatics, and Decision Systems Engineering. Before joining ASU, he was an associate professor of the School of Computing and Information Sciences (SCIS) at Florida International University. He directs the Research Laboratory for Virtualized Infrastructure, Systems, and Applications (VISA). His research interests are in distributed/cloud computing, big data, high-performance computing, autonomic computing, virtualization, storage systems and operating systems.