CS 595-xx - Topics in Modern Big Data Analytics

Course Description:

Big data technologies, in particular, scalable distributed platforms for storage and analytics enable processing of massive datasets for analytics, machine learning, and other use cases. This course provides a comprehensive overview of algorithms, systems, and techniques for Big Data processing. In a semester-long project, students will extend existing big data platforms. Additionally, in the seminar component of this course we will discuss cutting edge research and industrial developments in the field.

Course Material:

The following text book will be helpful for following the course and studying the presented material.

White, **Hadoop: The Definitive Guide**, 4th Edition, O'Reilly Media, 2015

One of the following standard text books on databases in general may be helpful. However, this is not required reading material.

Elmasri and Navathe. **Fundamentals of Database Systems**, 6th Edition , Addison-Wesley , 2003
Ramakrishnan and Gehrke. **Database Management Systems**, 3nd Edition , McGraw-Hill , 2002
Silberschatz, Korth, and Sudarshan. **Database System Concepts**, 6th Edition , McGraw Hill , 2010
Garcia-Molina, Ullman, and Widom. **Database Systems: The Complete Book**, 2nd Edition, Prentice Hall, 2008

Slides for the course will be made available on the course webpage.

Prerequisites:

No formal prerequisites, but some background in databases and/or distribute programming is useful.

Course Details:

The following topics will be covered in the course:

- **Foundations of Scalable and Distributed Storage and Computation**
  - Fault tolerance
  - Eventual consistency and consensus protocols
  - Load balancing
  - Scalable algorithm design
  - Data placement techniques
- **Distributed Storage**
  - Distributed file systems and replication
  - Key-value and distributed document Stores
  - Structured distributed storage solutions
- **Distributed Batch Processing**
  - Specifying computations as dataflows
  - DISC systems
  - Iterative and incremental dataflows
- **High-level Dataflow Languages**
  - Scripting and query languages
  - Graph processing
- **Streaming Analytics**
  - Distributed stream processing
  - Publish-subscribe systems
- **Distributed Transaction Processing**
  - The 2PC protocol
  - Transaction processing over partitioned storage

Workload

The workload will consist of

1. A semester long project related to extending an existing Big Data platform
2. Review a research paper related to state-of-the-art techniques in Big Data processing and present it in the course

Course Objectives:

After attending the course students should:

- Understand the challenges of processing queries and other data-intensive computations in a distributed fashion
- Be familiar with scalable storage and compute solutions; understand their benefits and limitations
- Learn about different types of scalable systems including ...
  - *Distributed file systems*
  - Scalable storage techniques such as *key-value stores* and distributed structured storage solutions such as *HBase*
  - DISC platforms such as MapReduce, Spark, and Flink
  - Specialized systems for, e.g., *graph data* such as Giraph and support for graph data in general purpose DISC platforms
  - *Publish-subscribe systems* such as Kafka
  - Distributed transaction processing systems
- Understand what *fault tolerance* is and how it can be achieved through replication, logical logging (as in Spark), and through *consensus protocols* like Paxos and Raft
- Understand how *load-balancing* is achieved in DISC systems
- Understand *data placement techniques* including horizontal and vertical partitioning and how they utilized by DISC frameworks
- Learn about the distributed algorithms employed by DISC platforms for implementing the higher-order functions exposed to the user

Grading Policy:

The grading scheme is as follows:

- A: 80% or higher
- B: 50% or higher
- C: 35% or higher

- E: below 35%

The weighting of the individual components are:

- Programming Project: 50%
- Literature Review: 50%

Illinois Tech's Sexual Harassment and Discrimination Information:

- Sexual harassment, sexual misconduct, and gender discrimination by any member of the Illinois Tech community is prohibited. This includes harassment among students, staff, or faculty. Sexual harassment by a faculty member or teaching assistant of a student over whom they have authority or by a supervisor of a member of the faculty or staff is particularly serious. Such conduct may easily create an intimidating, hostile, or offensive environment.
- Illinois Tech encourages anyone experiencing sexual harassment or sexual misconduct to speak with the Title IX Office for information on the resolution process and support options.
- You can file a complaint electronically at `http://iit.edu/incidentreport`, which may be completed anonymously. You may also file a complaint in-person by contacting the Title IX Coordinator, Virginia Foster at `312-567-5725` / `mailto:foster@iit.edu` or the Deputy Title IX Coordinator `312-567-5726` / `mailto:eespeland@iit.edu`.
- If you are not ready to file a formal complaint but wish to learn about your rights and options, you may contact Illinois Tech's Confidential Advisor service at `773-907-1062`. You can also contact a licensed practitioner in Illinois Tech's Student Health and Wellness Center at `312-567-7550`
- For a comprehensive list of resources regarding counseling services, medical assistance, legal assistance and visa and immigration services, you can visit the Title IX Office's website at `https://web.iit.edu/hea/resources`