# **Big Data (**2016-2018 **Fall)**

Instructor: Yinghui Wu

**Description**: This course introduces basic technology (algorithms, architectures, systems) and advanced research topics in connection with large-scale data management and information extraction techniques for big data. The course will start by introducing Big data models, databases and query languages, and cover modern distributed database systems and algorithms and Big data systems adopted in industry and science applications. Implementation of a distributed database on a standalone machine will be covered and students will learn how to build their own database for big data. Distributed storage and parallel processing and architectures that support data analytics will be examined, and students will learn how to implement a distributed data processing system. The course will also cover critical topics in mining and knowledge discovery of big data, with applications in social analytics, cyber security, and information networks, among others that are already in public eye.

1. Introduction:
   - Big Data: concept, research problems, hot research trends and emerging applications
   - Big Data models: Basics on relational data models and semi-structured data;
   - Graph Data: Graph data models; basics on graph theory
   - From Data to Knowledge: a knowledge discovery process.
   - **Course project**: choose your real-life datasets and justify a data model to present it.

2. From Data Models to Databases
   - Relational databases
   - Beyond Relational databases: noSQL systems, Key-value store, Column store, Big Tables
   - Graph databases: architecture, storage, indexing and graph views
   - **Course project:** choose and get familiar with an open source database system (provide list) and a query language

3. Extracting information from Big Data
   - Relational data querying: query languages and algorithms
   - Querying Big Data and Graphs: XML, SPARQL,
   - Feasible Big data Querying: Approximate querying models
   - Making Big Data small: querying optimizations, compression, sampling techniques
   - Drinking from a firehose: Querying dynamic graphs and data streams; incremental query evaluation; answering query using views
   - **Course project**: Based on your earlier implementation, test a "big data" search idea.

4. Scalable Big Data Processing: platforms and systems
   - Distributed and parallel computing models
   - Parallel querying, distributed processing models and systems
     - MapReduce, Apache Mahout and Giraph project;
     - vertex-centric models, GraphX, Pregel and GraphLab
   - **Course project:** Implement a parallel/distributed version of your earlier algorithm using MapReduce and cloud services e.g., Amazon EC2.

5. From Big Data to Big Knowledge (Course Project)
   - Data Mining over massive datasets
   - Storytelling: Big data visualization
   - Big data, Big ethics: Privacy and Security
   - **Course project**: Write up a scientific report on knowledge discovery process in your project.

**Prerequisites**:
Students are expected to have basic programming experiences and knowledge of algorithm design, equivalent to a data structure course. Some background in basic linear algebra is necessary. The course will cover basics in graph theory and relational databases.

**Grading:** The course consists of lectures (twice a week), 4 assignment and a small project running through the course. The course project leads to a complete implementation of a small analytic tool over a specific dataset. Students are encouraged to work in teams for the course project. A final project report will be given by each team. A reading list will be provided for offline learning. Final grades will be determined as follows:

- Class participation: 10%
- Homework: 40%
- Project: 40%
- Final Project report and presentation: 10%

Note: You may come up with your own project topics and proceed with the permission of the instructor. The proposed topic must be closely related with the course topics.

**References**:

- **Database Systems: The Complete Book**
  **https://www.amazon.com/Database-Systems-Complete-Book-2nd/dp/0131873253**
- **Big Data: A Revolution That Will Transform How We Live, Work, and Think** http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544227751
- **Mining of Massive Datasets**. Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. v2.1, Cambridge University Press. 2014. (free online)
- **Hadoop: the definitive guide** http://hadoopbook.com/
- **Graph Databases**
- http://www.goodreads.com/book/show/17465372-graph-databases?from_search=true&search_version=service

**Policies**

#### Missing or late work

Except by prior arrangement, missing or late work will be counted as a zero.

#### Collaboration

Exploratory collaboration on assignments is allowed (and the final submissions must contain a list of all collaborators). However, students must prepare solutions individually. As an example of the ideal scenario, the following situation is permissible:

A group of students meets to develop the solution to a problem on a white board. Each student records individual notes from this problem solving meeting. All students then prepare solutions individually and without further collaboration. These solutions show the names of all members in the initial group.

Examples of collaborations that are not allowed include, but are not limited to:

- Sharing pieces of any written solution
- Sharing source code or any other computer programs
- Reviewing final written solutions

Collaboration on the projects must be discussed with the instructor.