Database Design and Implementation
(Big Data Analytics)

Prof. Yanlei Diao
University of Massachusetts Amherst
Information Technology has been evolving rapidly
In 50 years, computers will be intelligent, with a storage capacity of about $10^9$.

--- Alan Turing, 1950

1950
Turing’s Vision

1970
Codd: Relational Model & DBMSs

1998
Gray: Turing Award
When processing, storage, and transmission cost micro-dollars, then the only real value is the data ... 

-- James Gray, 1998 Turing Award Speech
Era of BIG DATA

1950  Turing’s Vision
1970  Codd: Relational Model & DBMS
1998  Gray: Turing Award

New Millennium
Era of BIG DATA

- Relational algebra
- Data mining
- Deep learning
- Distributed algorithms
- Stream algorithms

- Sensors
- Smart devices
- Sequencers
- Large simulations
- Social software
- ...

New Millennium

1950
Turing's Vision

1970
Codd: Relational Model & DBMS

1998
Gray: Turing Award

2003
Cluster Computing

2006
Cloud Computing Hadoop

2009
Spark
Objectives of This Course

Fundamental **Design Principles** and **Key Algorithms** for Influential Software Systems for Big Data Analysis

![Diagram showing relationships between Algorithms, Data, and Infrastructure]
Main Topics

1. Relational Database Management Systems (RDBMS)
   Apps: banking, ticketing, electronic commerce…

- Data Modeling and Query Languages
  - Relational model, relational algebra/calculus, SQL, theory
- Query Processing
  - Algorithms for sorting, selection, join, group by-aggregation.
- Query optimization
  - System R style, multi-objective, parametric optimization
- Transaction processing: consistency and recovery
Main Topics (cont’d)

2. Data Warehouses, Data Mining, and Data Exploration

Apps: retail management, fraud detection, decision analytics…

- Online Analytic Processing (OLAP)
  - Large queries over (read-only) historical data
  - OLAP operations
  - New querying and optimization methods
- Data mining algorithms
- Data exploration
Main Topics (cont’d)

3. Parallel Databases & MapReduce Systems

Apps: retail management, web data analysis, social networks…

- **Parallel databases (90’s)**
  - Programming model: SQL
  - Data-parallel algorithms for relational operators

- **MapReduce systems (circa 2003-2004)**
  - Programming model: functional
  - MapReduce under data parallelism
  - Scheduling and fault tolerance

- **Distributed storage**
  - HDFS, key-value stores, column stores
Main Topics (cont’d)

4. Unified Analytics: SQL, Machine Learning, Graph, etc.

Apps: SQL, machine learning, graph analytics…

- Spark (2009-)
  - Programming model: RDDs and data frames
  - Innovation in fault tolerance
  - Unified programming interface for SQL, Machine Learning, graphs, and stream workloads
Putting it All Together

- SQL
- Scripting for ETL
- Stream Processing
- Machine Learning
- Graphs

Execution Engine (MapReduce, Spark, SQL)

Distributed Storage (file systems, key-value stores)
From the Business’ Perspective

Data sources

ETL (Spark / Hadoop)

Periodic / Ad-hoc Exploration (SQL)

Targeted Complex Analytics (ML)
Teaching Staff

- Instructor: Prof. Yanlei Diao
  - Email: yanlei@cs.umass.edu

- Teaching Assistants
  - Shivam Srivastava, shivam@cs.umass.edu
  - Matteo Brucato, matteo@cs.umass.edu

- Graders
  - Kajal Tiwari, kajaltiwari@umass.edu
  - Nitin Srinivasan, nitinsriniva@umass.edu
Meeting Time and Prerequisites

- **Tue/Thu, Feb 2 – May 4**
  - 8:30 – 9:45 am
  - By zoom invitation

- **Prerequisites**
  - Undergraduate data structures and algorithms
  - Undergraduate databases or Operating Systems
  - Optional: undergraduate course on Artificial Intelligence or Machine Learning
Course Website and Resources

- Course website
  
  http://avid.cs.umass.edu/courses/645/s2021

- Piazza for questions and answers
  
  https://piazza.com/umass/spring2021/cmpsci645/

- Gradescope for homework and project submissions

- Timezone questionnaire
Database Management Systems
3rd Edition
Ramakrishnan and Gehrke

Amazon:
- Buy new: $43-$147.09 (hardcover); paperback, $23; Kindle, rent options are also available…

Lecture notes and recorded videos will be posted on the course website after class.
Course Workloads

- Class participation (5%)
  - 20 / 27 registered lectures
  - Missed lectures can be made up by answering questions in Piazza

- Six homework assignments (50%)
  - Written problem sets
  - Programming exercises using PostgreSQL and Spark

- Midterm exam (20%)

- Reproducibility project in teams of three (25%)
  - Read a paper, write a review, implement techniques, reproduce results (e.g., plots, tables) on a given dataset

- Optional: Open-ended research project = reproducibility project + homework 5 + homework 6
  - Prerequisite: prior research, a PhD student or having a publication
  - Expected to design new techniques with substantial results
Interaction between Staff and Students

- Online lectures that require registration
- Lecture notes and recorded videos will be posted on the course website
- Office hours of the 645 staff that span Mon – Fri
  - Questions regarding the homework, exam preparation, grading, and project
- Questions can be posted on Piazza and will be answered on a daily basis, Mon – Fri
Academic Honesty

- All submitted work must be your own!
  - Although students are encouraged to study together, each student must produce his or her own solution to each homework.
    - Copying or using sections of someone else’s program or assignment (even if it has been modified by you), or copying a solution from an external source, is not acceptable.
      - The teaching staff will be vigorous in enforcing them.