www.cdf.toronto.edu/~csc343h/fall

CSC 343
Introduction to Databases

Diane Horton and Renée Miller
Fall 2016

UNIVERSITY OF
TORONTO

Our first hour or so

- Some key concepts
- Examples to motivate the course
- Admin info

Databases and DBMSs

- Databases are everywhere, often behind the scenes.
- DBMS (Database Management System):
  "A powerful tool for creating and managing large amounts of data efficiently and allowing it to persist over long periods of time, safely." [Ullman and Widom, FCDB]
- Database:
  a collection of data managed by a DBMS.

Data models

- Every DBMS is based on some data model: a notation for describing data, including
  - the structure of the data
  - constraints on the content of the data
  - operations on the data
- Some specific data models:
  - network & hierarchical data models — of historic interest
  - relational data model
  - semistructured data model
  - unstructured data — (key, value) pairs
    - value could be anything, even a full document

The relational data model

- Main concept is a "relation."
  Based on the concept of relations in math.
- Can think of as tables of rows and columns.

Teams

Name	Home Field	Coach
Rangers	Runnymede CI	Tarvo Sinervo
Ducks	Humber Public	Tracy Zheng
Choppers	High Park	Ammar Jalali

Games

Home team	Away team	Home goals	Away goals
Rangers	Ducks	3	0
Ducks	Choppers	1	1
Rangers	Choppers	4	2
Choppers	Ducks	0	5

Example …

- A dataset scraped from Twitter
- Defining a schema that expresses its structure
- Creating an instance that contains the data
- Writing some queries on the data

What a DBMS provides

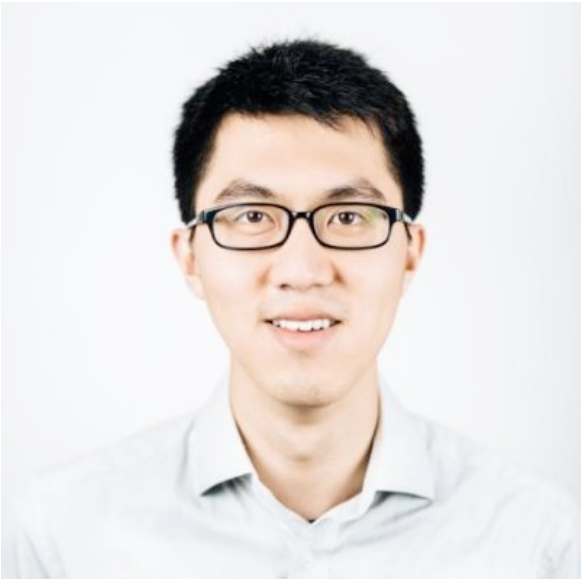
- Ability to specify the logical structure of the data
  - explicitly
  - and have it enforced
- Ability to query or modify the data.
- Good performance under heavy loads (huge data, many queries).
- Durability of the data.
- Concurrent access by multiple users/processes.

Overall architecture of a DBMS

- The DBMS sits between the data and the users or between the data and an application program
- Within the DBMS are layers of software for:
  - parsing “queries”
  - implementing the fundamental operations
  - optimizing queries
  - maintaining indices on the data
  - accessing the files that store the data and indices
  - management of buffers
  - management of disk space

Big Data Architectures

- Provide scale-out to hundreds of machines
- Provide fault-tolerance if some machines fail
- Originally designed for unstructured data
  - Today we are seeing DBMS and Big Data functionality merging in Big Data Analysis platforms like Spark
- Historical note: Reynold Xin (co-founder of Databricks — Apache Spark) is a UofT grad.



What this course is about

- csc443 is about implementation of the DBMS itself
- csc343 is about *using* DBMSs:
  - defining schemas and instances
  - writing queries
  - connecting to code written in a general-purpose language
  - rigorous underlying principles

Why study databases?

- Interesting concepts and techniques.
- Spans computer science, including OS, languages, theory, AI, multimedia, logic.
- Databases have become increasingly important
  - shift from a focus on computation to information
  - data increases in volume and diversity.
- Jobs: In demand and well paid.
- Research: Many open problems.