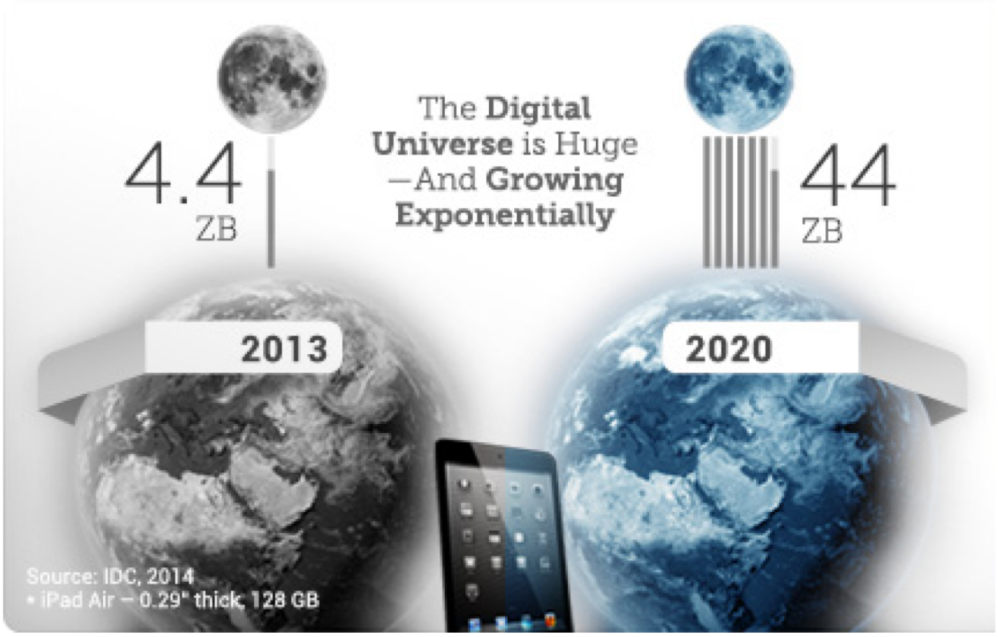
CMPT 354: Database System I

Lecture 1. Course Introduction

Outline

- Motivation for studying this course
- Course admin and set up
- Overview of course topics

Trend 1: Data grows exponentially



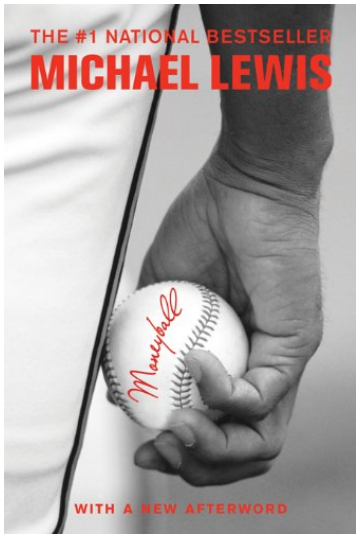
1 ZB = 1, 000, 000, 000, 000 GB

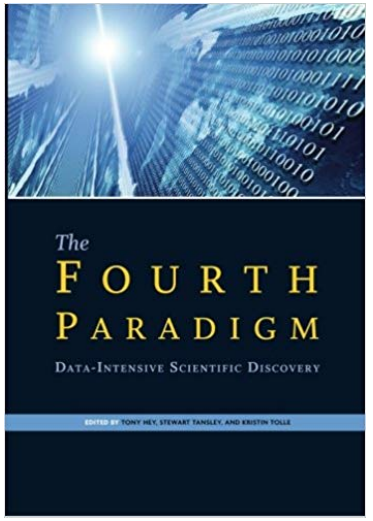
Why Trend 1?

- Human Generated Data
  - Social Media
  - Camera/Microphones
  - Activity Trackers
  - …
- Machine Generated Data
  - Software Logs
  - Smart Home
  - Self-driving Car
  - …

Trend 2: Data skills are in increasingly high demand

Why Trend 2?





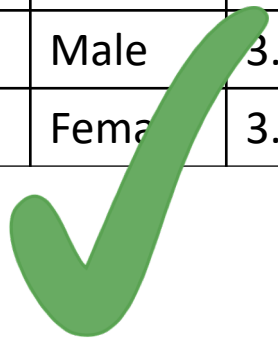
everything

Data is at the center of ~~many things~~

Database

- What is a database?
  - A collection of files that store related data
- We will mainly focus on **relational** databases (i.e., data is stored in tables)

Name	Gender	GPA
Mike	Male	4.0
Bob	Male	3.6
Alice	Fema	3.8



Text Data

Photos & Videos

Databases in Real Life

- Examples
  - Amazon: Online Bookstore
  - SFU: Course Management System
  - RBC: Banking System
  - Air Canada: Airline Reservation System

- Answer two questions
  - What data do they need?
  - What applications do they need to build?

Amazon: Online Bookstore

- Database
  - Data about books, customers, orders, etc.
  - Data about sessions (clicks, pages, searches)

- Applications
  - Book Search System
  - Recommender System
  - Payment System
  - Order System
  - ….

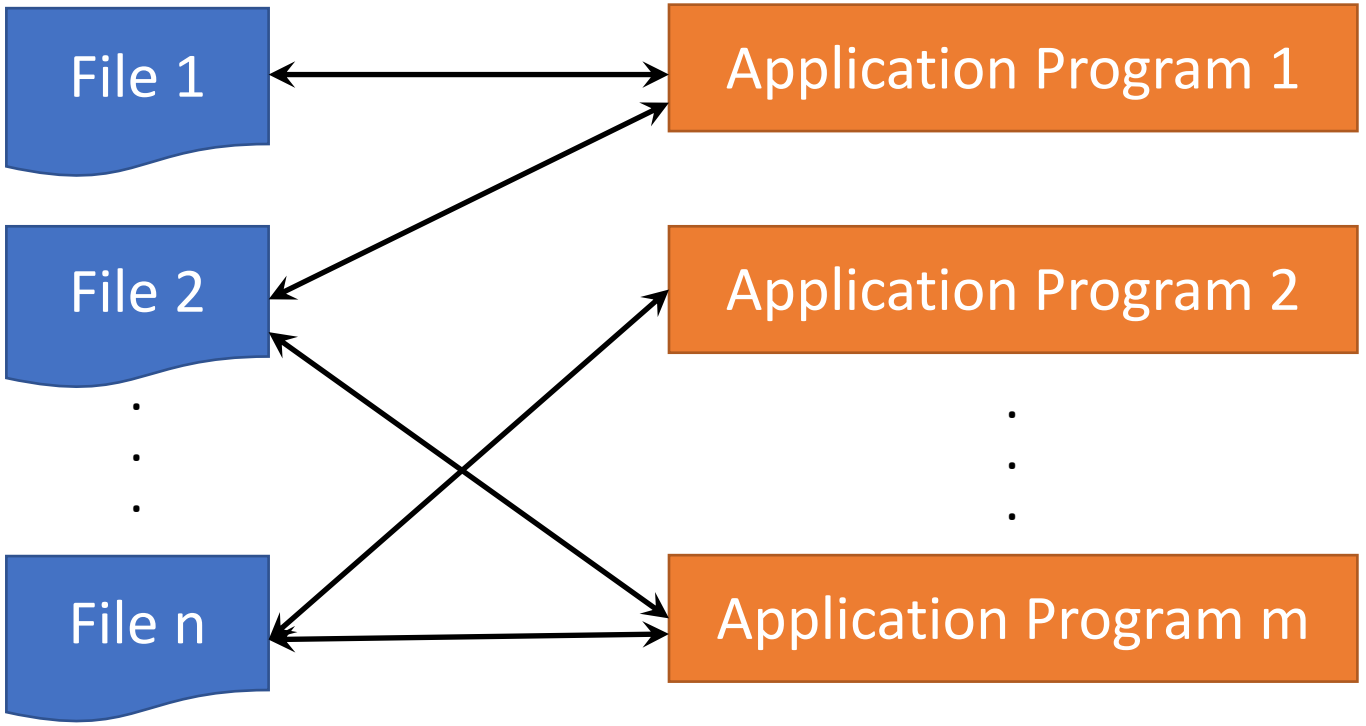
Database Management Systems (DBMSs)

- What is a DBMS?
  - A piece of software designed to store and manage databases

- Examples
  - Commercial: Oracle, IBM DB2, Microsoft SQL Server
  - Open source: MySQL (Sun/Oracle), PostgreSQL, SQLite

Data Storage without DBMS

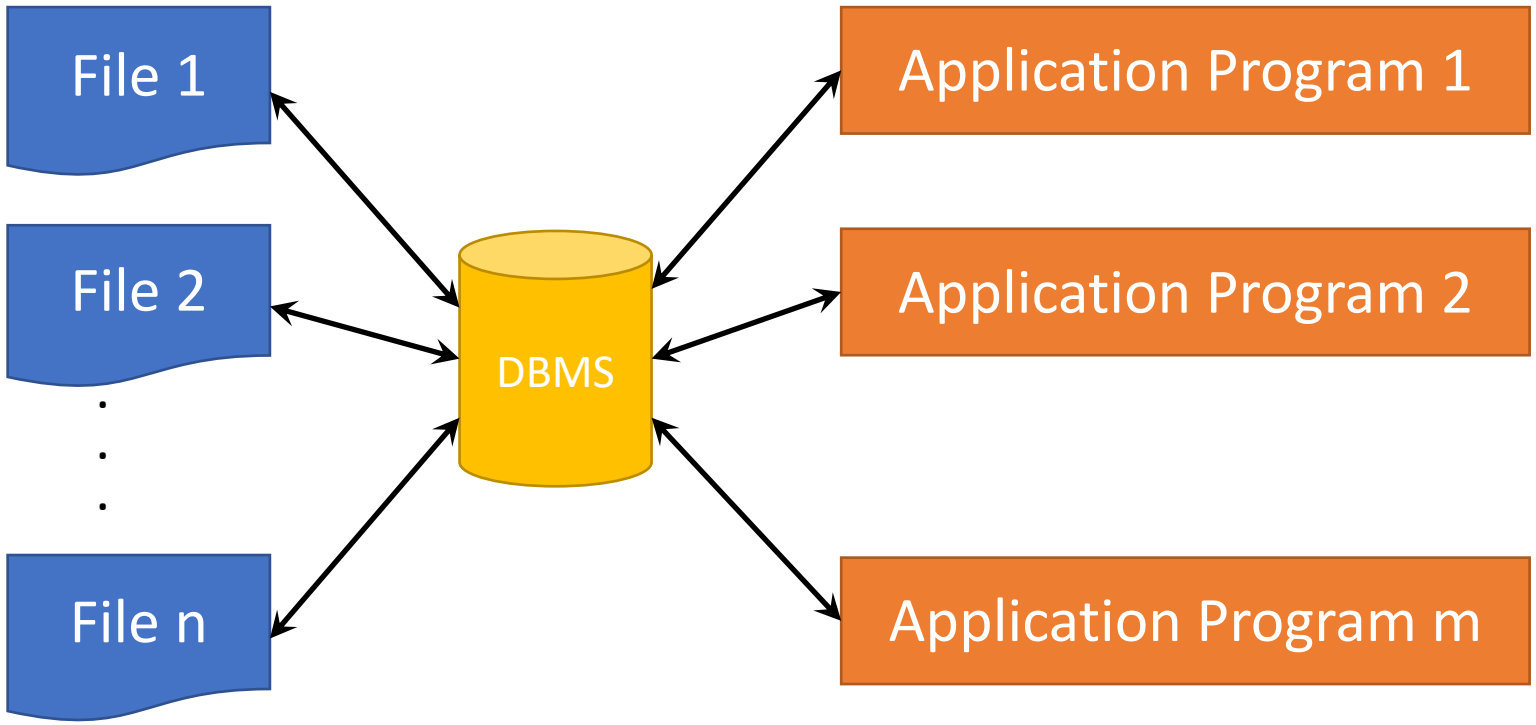
- Data would be collected in many different files and
- Used by many application programs



What happens if

- Several programs need to access and modify the same record at the same time?
- An attribute is added to one of the files?
- We need to repeatedly access a single record out of millions of records?
- We need to retrieve data stored in multiple files?
- The system crashes while one of the application programs is running?

Data Storage with DBMS



DBMS Functions

- All access to data is centralized and managed by the DBMS
- Use advantages
  - Efficient access
  - Data integrity and security
  - Concurrent access and concurrency control
  - Crash recovery
- Design and implementation advantages
  - Logical data independence
  - Physical data independence
  - Reduced application development time

Current Market

- Relational database still anchor the software industry
  - Elephants: Oracle, IBM, Microsoft, Teradata, EMC, …
  - Open source: MySQL, PostgreSQL
  - Emerging variants: In-memory, Column-oriented
- Open source “NoSQL” is growing
  - Analytics: Hadoop MapReduce, Spark
  - Key-value Stores: Mongo, Cassandra, Couch
- Cloud services are expanding quickly
  - Amazon Redshift/Aurora, Microsoft Cosmos DB

Course Objectives

1. Master skills to **query a *database***
2. Master skills to **design a *database***
3. Understand **how a *DBMS* works**

Who needs this course?

- **DB designer:** establishes schema
- **DB application developer:** writes programs that query and modify a database
- **DB administrator:** tunes systems and keeps whole things running

- **Data scientist:** manipulates data to extract insights
- **Data engineer:** builds a data-processing pipeline

Outline

- Motivation for studying DBs
- **Course admin and set up**
- Overview of course topics

Staff

- Instructor:
  - Jiannan Wang (jnwang@sfu.ca)
  - Faculty (joined SFU in 2016)
  - Office hours: Wednesday 10:30-12:00 (noon), TASC 1 9237
- TA:
  - Changbo Qu (changboq@sfu.ca)
  - PhD student (joined SFU in 2017)
  - Office hours: Tuesday 10:30-12:00 (noon), TASC 1 9217

Course Format

- Lectures
  - Location: AQ3149
  - **PLEASE ATTEND!**
- Five homework assignments
- Midterm and final

Grading

- Homework: 5 * 6% = 30%
- Midterm: 30%
- Final: 40%
- This is all subject to change

Communications

- Web page
  - Link: https://sfu-db.github.io/cmpt354
  - Course information, lecture notes, and assignments
- Piazza
  - Sign up: https://piazza.com/sfu.ca/fall2018/cmpt354
  - THE place to ask course-related questions
  - Log in today and enable notifications
- Class mailing list
  - You are automatically subscribed
  - Low traffic, only important announcements
- Google form
  - Link: https://goo.gl/forms/UH0nvxKGAFNMkCtr1
  - Provide anonymous feedback to improve courses

Textbooks

- **[GUW]** Database Systems: The Complete Book (2nd Edition)
  - Hector Garcia-Molina,
  - Jeffrey Ullman,
  - Jennifer Widom
- **[RG]** Database Management Systems (optional)
  - Raghu Ramakrishnan
  - Johannes Gehrke

Five Assignments

- A1. Basic SQL Queries
- A2. Advanced SQL Queries
- A3. Relational Algebra & Indexing
- A4. Schema Design
- A5. Transactional Application

Policy

- **Don’t be late**
  - You have up to 4 late days
  - No more than 2 on any one assignment
  - Once it is used up, 20% per day for each late day

- **Don’t Cheat**
  - We will do plagiarism check at the end of semester
  - If you got caught, your final mark would be deducted by 30%

Outline

- Motivation for studying DBs
- Course admin and set up
- **Overview of course topics**

CMPT 354 Topics

- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop
- **Week 15.** Final Exam

CMPT 354 and 454

- CMPT 354
  - How to query a database
  - How to design a database
  - How DBMSs work (basics)

- CMPT 454
  - How DBMSs work (advance)
  - How to implement DBMSs

Why should you care?

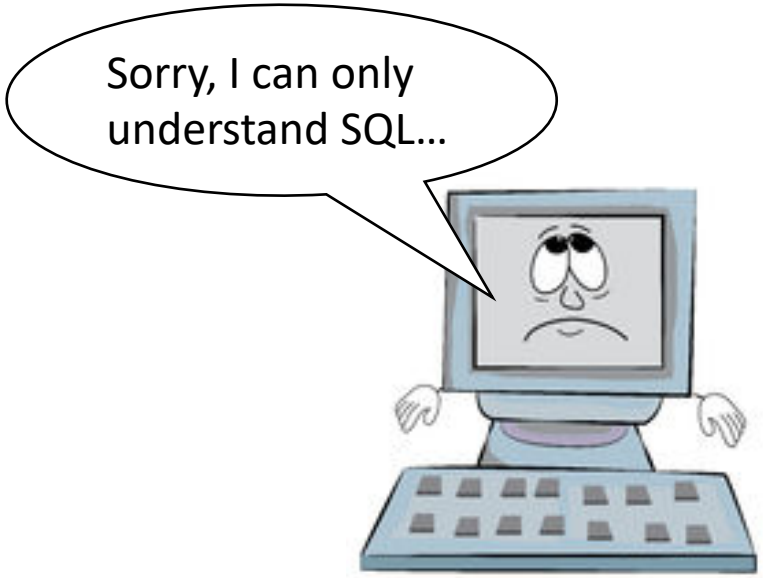
- **Week 2. Relational Model**
  - Ted Codd won a Turing Award by proposing the relational model
  - 5 out of 6 top database engines are relational databases

Rank			DBMS	Database Model	S
Aug 2018	Jul 2018	Aug 2017			Aug 2018
1.	1.	1.	Oracle	Relational DBMS	1312.02
2.	2.	2.	MySQL	Relational DBMS	1206.81
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1072.65
4.	4.	4.	PostgreSQL	Relational DBMS	417.50
5.	5.	5.	MongoDB	Document store	350.98
6.	6.	6.	DB2	Relational DBMS	181.84

- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop

Why should you care?

- **Week 3-4. Structured Query Language (SQL)**
  - Enable you to communicate with a DBMS
  - Declarative language (i.e., say what you want not how to do it)



DBMS

Find names of all students with GPA > 3.5

```
SELECT name
FROM Student
WHERE GPA > 3.5
```

- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop

Why should you care?

- **Week 5. Relational Algebra**
  - SQL: What you want
  - Relational Algebra: How to get it

Find names of all students with GPA > 3.5

```
SELECT name
FROM Student
WHERE gpa > 3.5
```

$$\Pi_{name}(\sigma_{gpa>3.5}(Students))$$

- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop

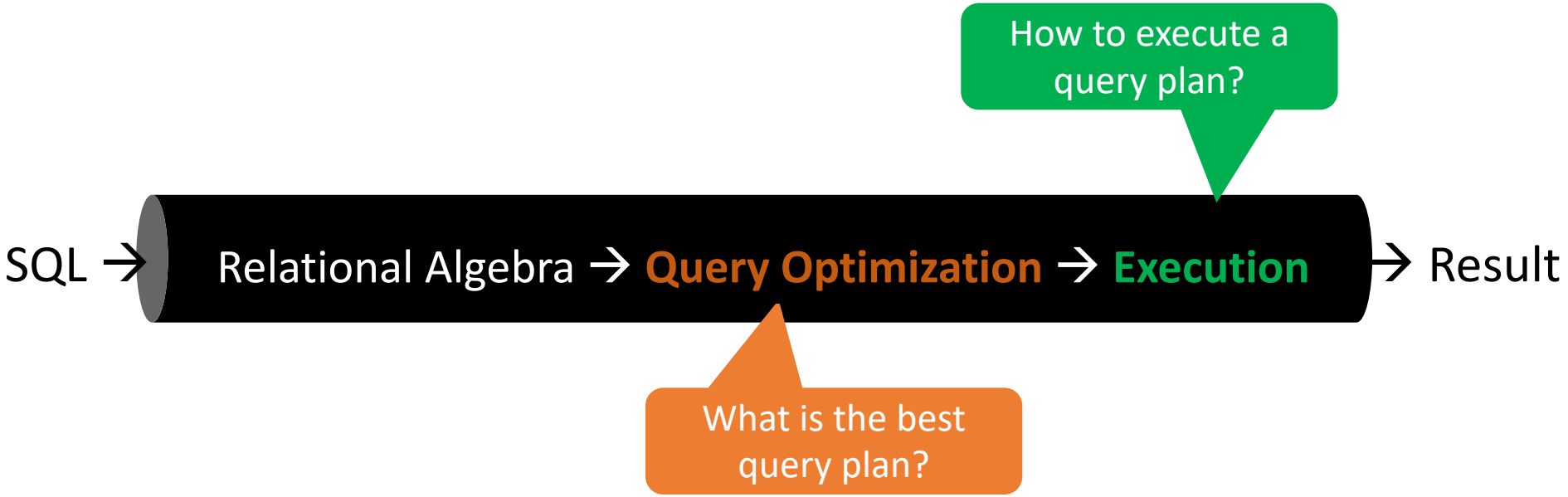
Why should you care?

- **Week 6. Storage and Indexing**
  - My database application is too **slow** … Why?
  - One of the queries is very **slow** … Why?

- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop

Why should you care?

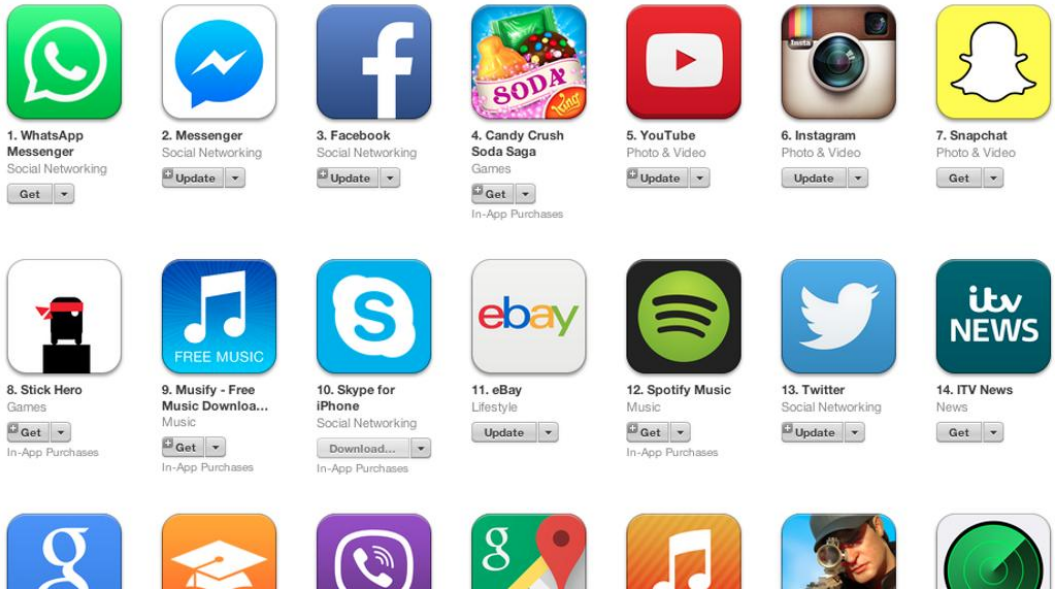
- **Week 8. Query Optimization and Execution**
  - Understand how an SQL query is processed



- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop

Why should you care?

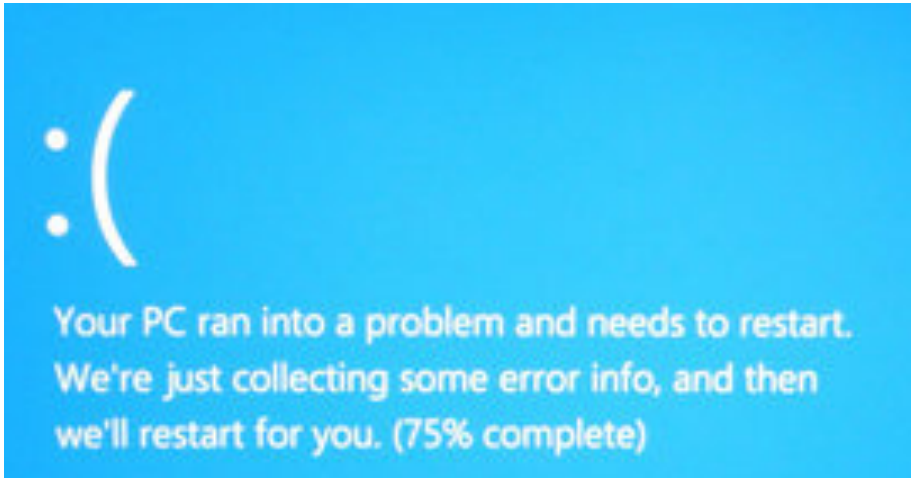
- **Week 9-11. Database Design**
  - How to design a database for an application (e.g. an iPhone APP)



- **Week 1.** Introduction
- **Week 2.** Relational Data Model
- **Week 3-4.** SQL
- **Week 5.** Relational Algebra
- **Week 6.** Data Storage and Indexing
- **Week 7.** Midterm
- **Week 8.** Query Processing
- **Week 9-11.** Database Design
- **Week 12.** Transaction Processing
- **Week 13.** NoSQL & SQL over Hadoop

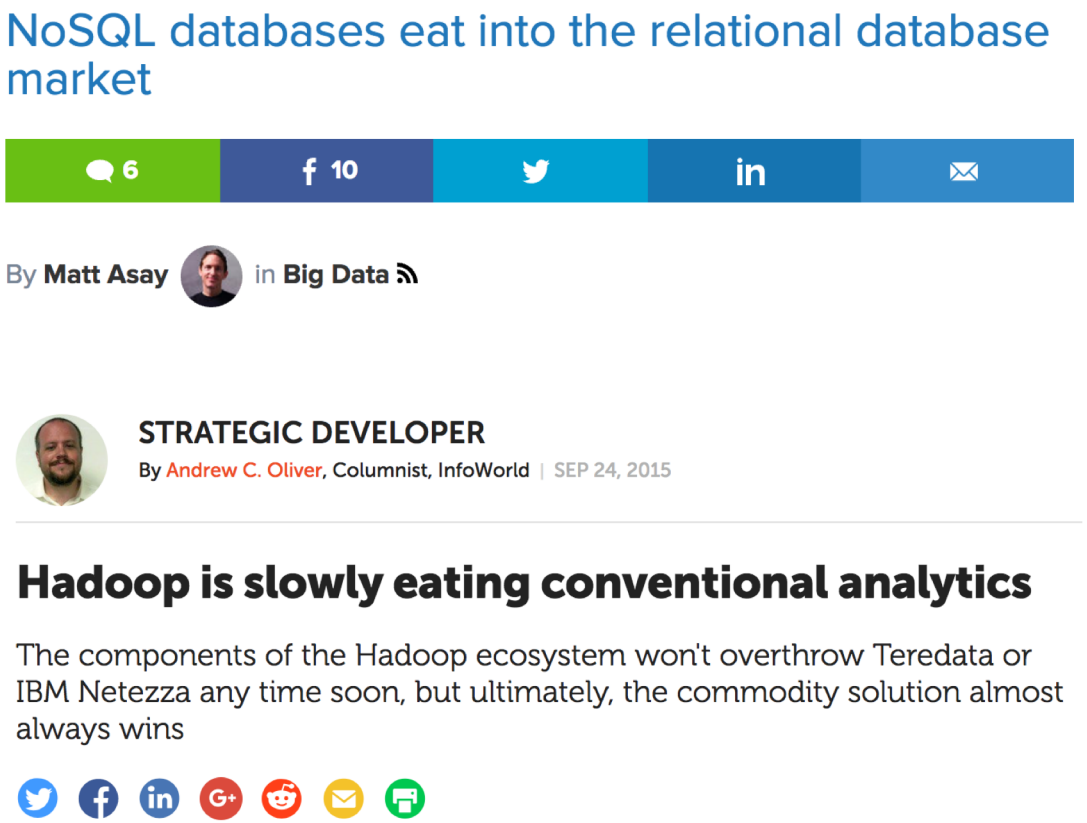
Why should you care?

- **Week 12. Transaction Processing**
  - What if multiple users access the same data
  - What if your computer crashes



Why should you care?

- **Week 13. NoSQL & SQL over Hadoop**



NoSQL databases eat into the relational database market

6 | 10 | in

By Matt Asay in Big Data

STRATEGIC DEVELOPER

By Andrew C. Oliver, Columnist, InfoWorld | SEP 24, 2015

Hadoop is slowly eating conventional analytics

The components of the Hadoop ecosystem won't overthrow Teredata or IBM Netezza any time soon, but ultimately, the commodity solution almost always wins

What to do next?

- Decide whether this is the right course for you

- Sign up Piazza and enable notifications
  - https://piazza.com/sfu.ca/fall2018/cmpt354

- Check out the course website
  - https://sfu-db.github.io/cmpt354/

Acknowledge

- Some lecture slides were copied from or inspired by the following course materials
  - “W4111: Introduction to databases” by Eugene Wu at Columbia University
  - “CSE344: Introduction to Data Management” by Dan Suciu at University of Washington
  - “CMPT354: Database System I” by John Edgar at Simon Fraser University
  - “CS186: Introduction to Database Systems” by Joe Hellerstein at UC Berkeley
  - “CS145: Introduction to Databases” by Peter Bailis at Stanford
  - “CS 348: Introduction to Database Management” by Grant Weddell at University of Waterloo