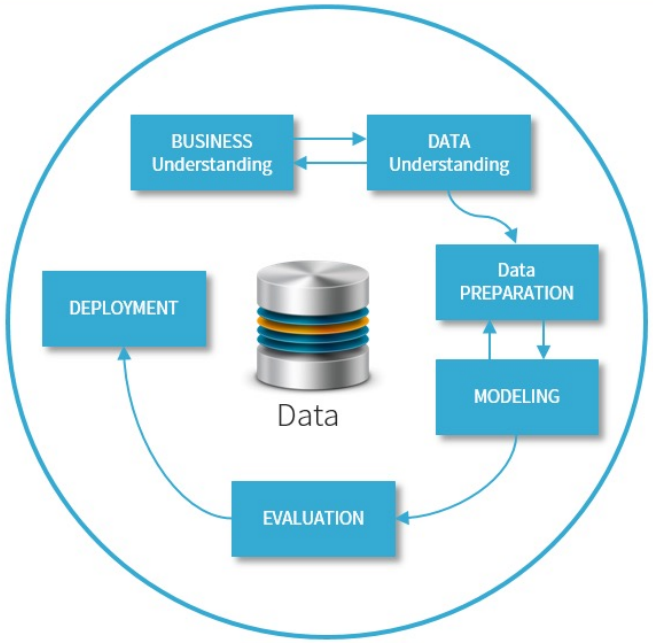
Introduction to Database Systems

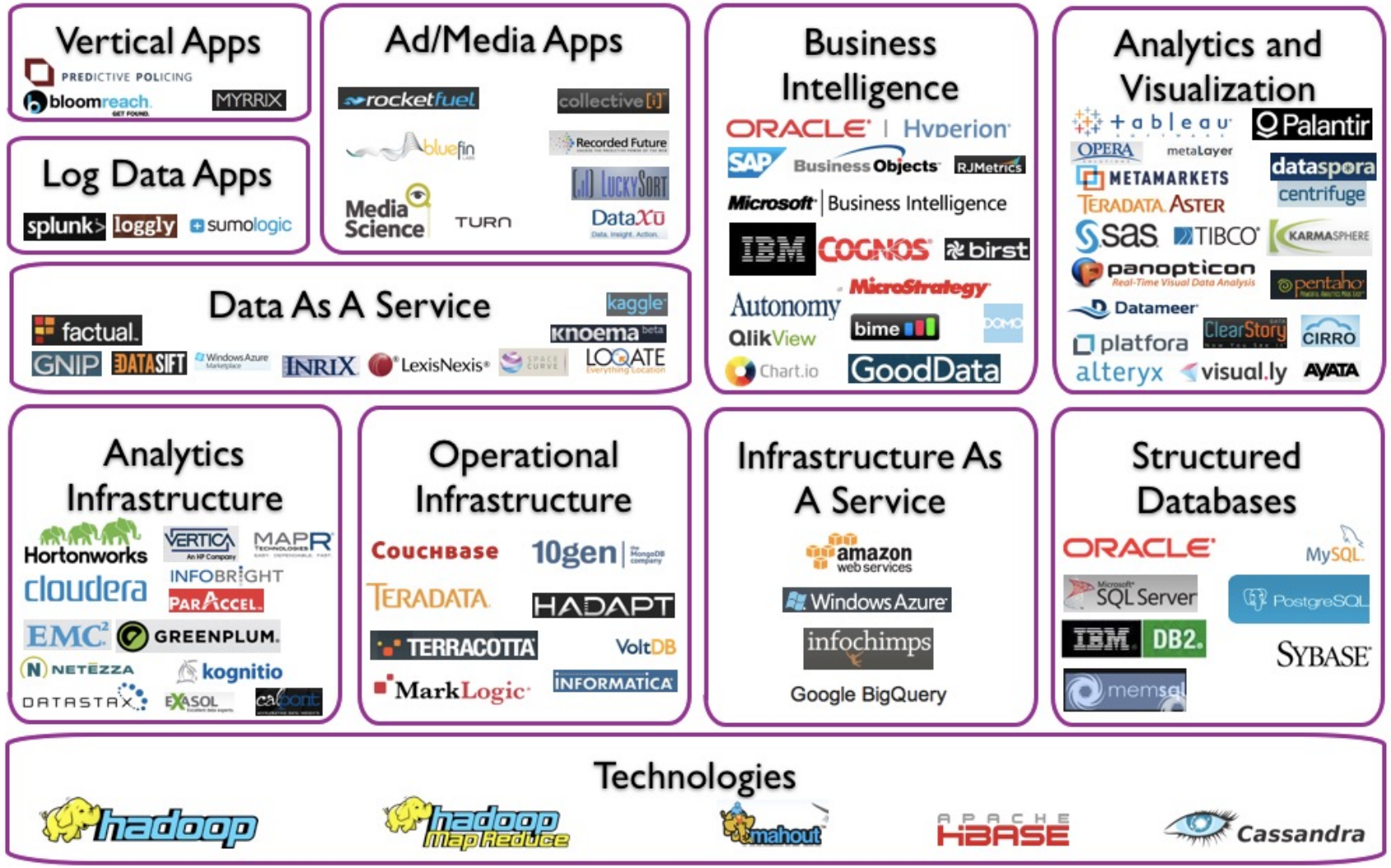
2023-Fall

About me

- B.A. in Computer Science and Technology, XJTU (4 years)
- Ph.D. in Software Engineering, THU. (6 Years)
- Lecturer in SEU (1 year)
  - Data mining, Deep Learning, Sequence Modeling, Internet of Things.

Big Data Ecosystem





Resources

CS w186
Introduction to Database Systems

Prof. Joe Hellerstein

Berkeley
cs186

CMU 15-445/645

CARNEGIE MELLON
DATABASE GROUP

https://db.cs.cmu.edu

- **Textbook:**
- Nengbin Wang, “Textbook of Database Systems”
- Database Management Systems,
- Database System Concepts (7th Edition)
- Database Systems: The Complete Book

Essential Queries

- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

Why-R1: Utility

- This class is very, very useful
  - Data processing backs essentially every app
  - Databases of one form or another back most apps
  - The *principles* taught in this class back nearly everything in computing

精选榜单

必吃榜 口味至上

必住榜 度假特色

必玩榜 城市打卡

今日排行 4月15日已更新 评选规则 >

美食 休闲娱乐 景点/周边游 酒店

全部美食 火锅 日本菜 西餐 本帮江浙

全城 热门榜 好评榜 口味榜 环境榜 服

TOP 01 [redacted]饭店·LAB...

4.7 ¥82/人

青年路/常青路 茶餐厅

推荐理由: 特别推荐, 惠灵顿牛排:鲜、嫩、爽口

TOP 02 [redacted]（[redacted]店）

4.8 ¥171/人

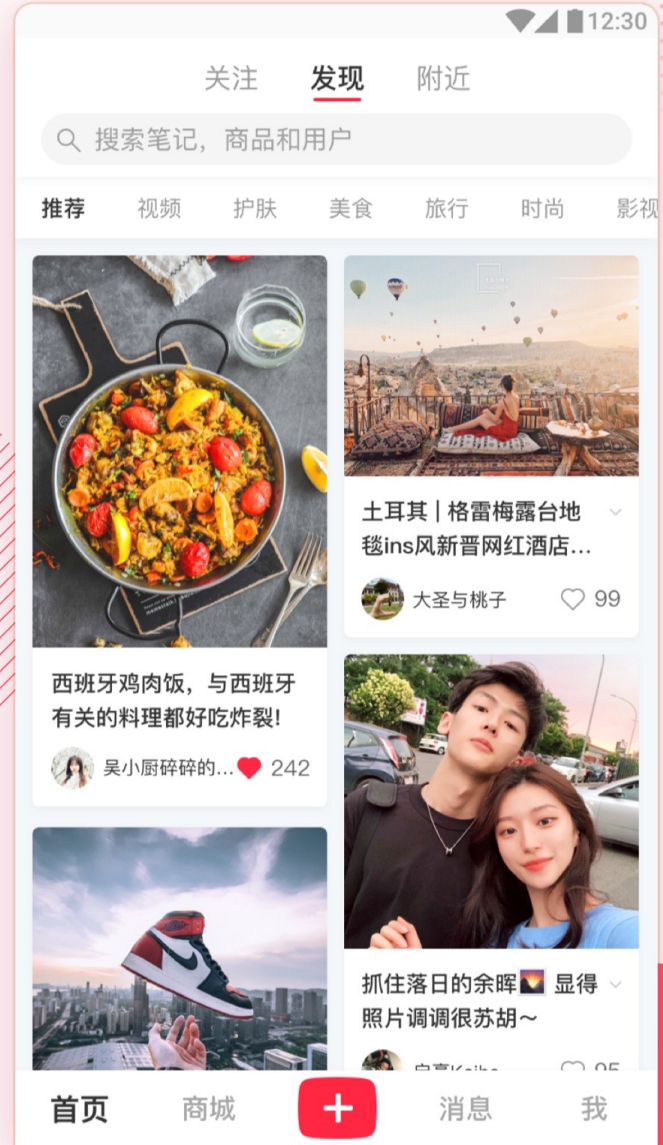
武展商圈 牛排

推荐理由: 店里的旅行主题设计风格,让人轻松愉悦

找答案

分享知识经验见解

吃穿玩乐买的日常



Why-R1: Utility

- The **fundamentals** of this class are (and will remain) central to participating in this new and more data-centric world
  - Many of the details and technologies will change in the coming years
    - Be prepared to generalize from what you learn here
    - Keep learning new things
- This material will empower you.

Why-R2: Centrality

- Data is at the center of modern society
- Data is unique in its nature and significance
  - *Particular* and *voluminous*
  - Often asymmetric
    - low value in isolation, high value when aggregated
  - Difficult to protect

Privacy Security Misinformation

USA TODAY

Home | News | Travel | Money | Sports | Life | Tech

Washington/Politics | Inside News

NSA has massive database of Americans' phone calls

Updat

Home / News & Blogs / IT Project Failures

Scathing report slams UK gov't data loss

By Michael Krigsman | July 1, 2008, 7:33am PDT

Summary:

guardian.co.uk

News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment

Money > Identity fraud

Zurich loses personal details of 51,000 customers

insurance firm says the data was lost during a routine transfer to South Africa in August last year, but there ... any misuse

Press Association, Thursday 22 October 2009 15.03

high 54. Weather map, Page B14.

Sky News Online

...king giant HSBC has apologised after revealing that details of 15,000 wealthy customers who held accounts with its Swiss private bank were stolen by ...

BBC

Home | News | Sport | Weather | TV | Radio

ONE-MINUTE WORLD NEWS

Timeline: Child benefits records loss

Two CDs containing personal details of 25m people have been lost by HM Revenue and Customs. Here is how the crisis unfolded.

The New York Times

© 2017 The New York Times Company | NEW YORK, FRIDAY, JANUARY 20, 2017 | $2.50

WIRETAPPED DATA USED IN INQUIRY OF TRUMP AIDES

TRUMP ARRIVES, SET TO ASSUME POWER

In Cabinet Hearings, Strong Rejection of Obama's Policies

The Sunday Teleg

Facebook was warned of data risks 7 years ago

By James Titcomb

said the company had no way of know-

Winning feeling

DOW JONES, A NEWS CORP COMPANY

DJIA ▲ 17873.22 0.25% | Nasdaq ▲ 4933.50 0.65% | U.S. 10 Yr ▼ 0/32 Yield 1.851% | Crude Oil ▲ 49.60 0.55% | Euro ▲ 1.1139 0.22%

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

SPECIAL OFFER: JOIN NOW

Home | World | U.S. | Politics | Economy | Business | Tech | Markets | Opinion | Arts | Life | Real Estate

New U.S. Study Fans Cellphone Cancer Worries

Samsung Adds More Ads to Its TVs

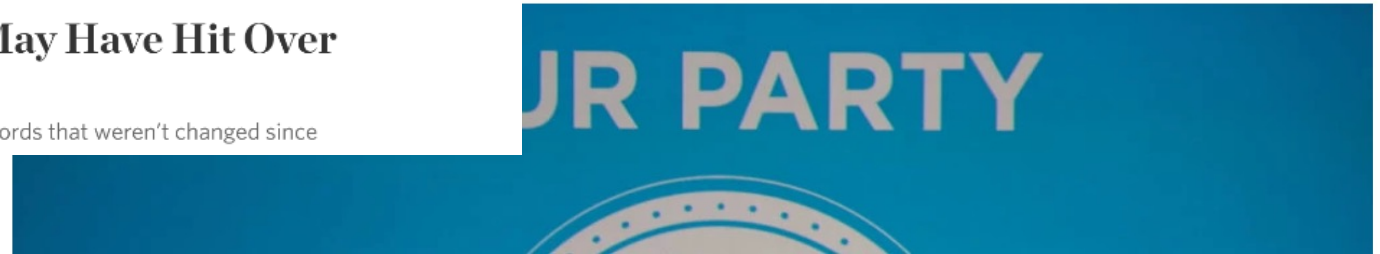
Verizon Workers Win Concessions in Deal to End Strike

TECH

LinkedIn 2012 Data Breach May Have Hit Over 100 Million

Professional social network says it will invalidate passwords that weren't changed since

...ty Says It Has Thwarted ...e of Voter Database



The Guardian
US edition

Opinion Facebook

Are you ready? Here is all the data Facebook and Google have on you

Dylan Curran

Google knows where you've been

Google knows everything you've ever searched – and deleted

Google has an advertisement profile of you

Google knows all the apps you use

Google has all of your YouTube history

Facebook stores everything from your s

location

They can access your webcam and micr

Google knows which events you attend

Google can know your workout routine

And they have years' worth of photos

Google has every email you ever sent

> "Manage to gain access to someone's Google account? Perfect, you have a diary of everything that person has done

How Much Data Does The NSA Look At Daily

TOTAL INTERNET TRAFFIC PER DAY

The NSA looks at **1.6 %** of the total Internet traffic, which is about **29 petabytes a day!**

(= 1 petabytes or 1,048,576 gigabytes)

For context, Google in 2010 said it had indexed only **0.004%** of the data on the net

...so, by inference from the percentages

Daily NSA DATA Collection = g+400

400 Googles?

(Based On 2010 Google Index Data)

Seven petabytes of photos are added to facebook each month. That's **.23 petabytes per day**

...so that means the NSA is 126 Facebooks

Daily NSA DATA Collection = 126 facebook

(Based On Uploaded Photos Data)

The master Netflix catalog takes up about **3.14 petabytes** of cloud storage space

Daily NSA DATA Collection = x 6,469,937 (4.7 gb DVD)

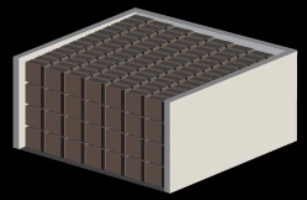
Just how much is 29 petabytes?

If a **3MB** smartphone photo is printed **8.5 inches wide**, the assembled **29 petabyte worth of photos** placed side by side would be over **1,392,000 miles long** - almost long enough to **wrap around the equator 60 times**

Enough to store the DNA of the entire population of the US – and then clone them, **58 times.**

x 21,623,967,288 (1.44 mb) floppy disks

x 3,801,088 Flash Drives



x 14,500 3000 sq.ft house w/ 8ft. ceilings, full of boxes of 3.5" floppy disks

Other facts

Google offers its users over **20 petabytes** (21.5 billion megabytes) of imagery — from satellite images to aerial photos to 360-degree Street View images

This is equal to the **whole production of hard-disk drives** in 1995

www.vsod.net

RESOURCES:

http://www.computerweekly.com/feature/What-does-a-petabyte-look-like

http://gizmodo.com/how-netflix-makes-3-14-petabytes-of-video-feel-like-it-498566450

http://www.theguardian.com/commentisfree/2013/aug/13/nsa-internet-traffic-surveillance

http://www.officedepot.com/a/products/438226/Imation-35-Diskettes-IBM-Format-DSHD/?cm_mmc=PLA-_-Google-_-Data_Storage_and_Media-_-438226-IQ_PE%20&Channel=Google&mr:trackingCode=01720179-71F3-DF11-8B7D-001B9BC043E8&mr:referralID=NA&mr:device=c&mr:adType=pla&mr:ad=28889904956&mr:keyword=&mr:match=&mr:filter=50816818436

Not all Data is Correct

"Any user can change any entry, and if enough users agree with them, ***it becomes true.****"*
– Colbert Report 7/31/2007

Asked viewer to update the page on Elephants to reflect a tripling population, forcing Wikipedia to lock the page.

COMEDY CENTRAL VIDEO ARCHIVE VIA WIKIPEDIA

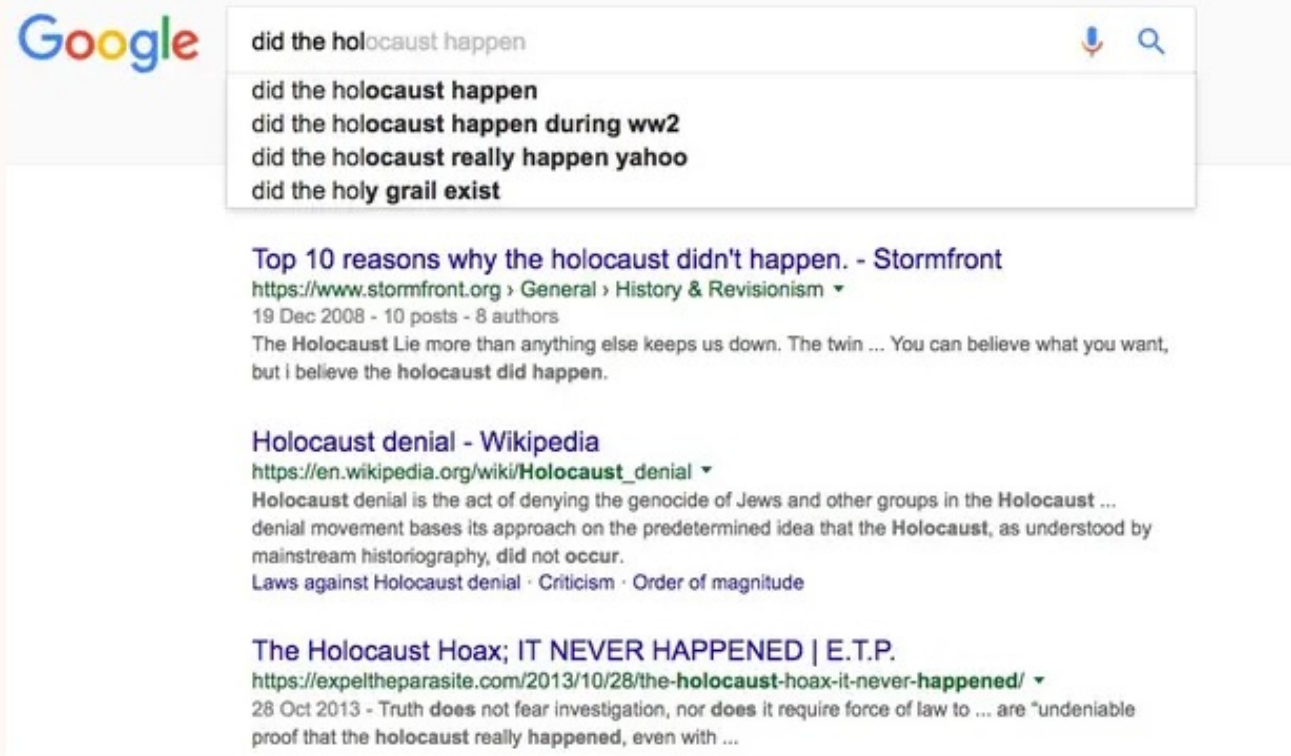
http://www.cc.com/video-clips/z1aahs/the-colbert-report-the-word---wikiality

Yet a 2005 Nature study found **Wikipedia science** articles to be ***similar in accuracy*** to ***Encyclopedia Britannica.***

https://en.wikipedia.org/wiki/Reliability_of_Wikipedia
http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

Not all Data is Correct

Last week, we reported how extremist sites 'game' the search engine, boosting their propaganda. In response, the web giant appears to have modified some results, but would like us not to notice



▲ Did the holocaust happen? Google search for Carole Cadwalladr Photograph: Google

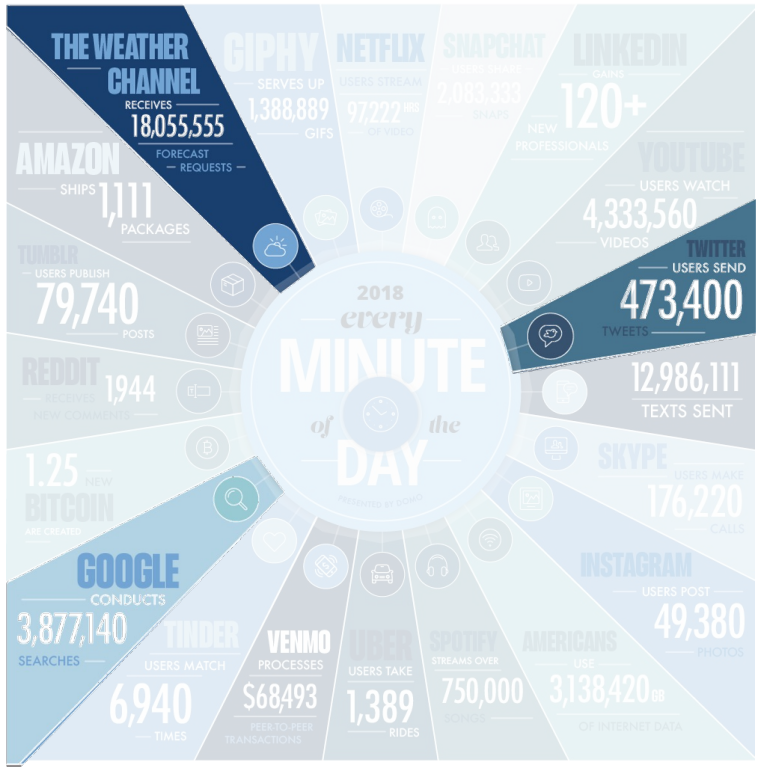
Google is not 'just' a platform. It frames, shapes and distorts how we see the world

Carole Cadwalladr

(From the Guardian, Dec 2016)

Why-R3: The Core of Computing

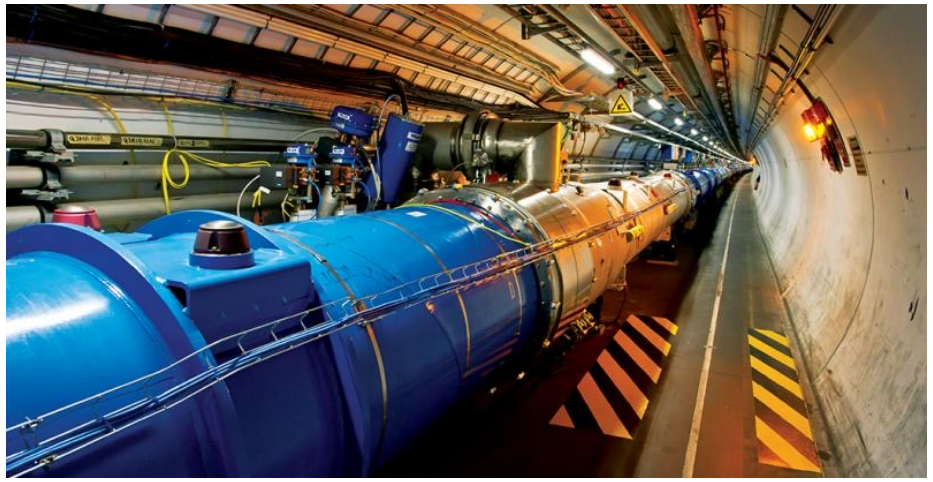
- Data growth will continue to outpace computation
- Systems for Data at Scale: the core of modern computing



Every Minute!

https://www.domo.com/learn/data-never-sleeps-5

Scale of Scientific Data



Large Hadron Collider, CERN

Metric prefixes in everyday use			
Text	**Symbol**	**Factor**	**Power**
yotta	Y	1 000 000 000 000 000 000 000 000	10^{24}
zetta	Z	1 000 000 000 000 000 000 000	10^{21}
exa	E	1 000 000 000 000 000 000	10^{18}
peta	P	1 000 000 000 000 000	10^{15}
tera	T	1 000 000 000 000	10^{12}
giga	G	1 000 000 000	10^{9}
mega	M	1 000 000	10^{6}
kilo	k	1 000	10^{3}

- Raw data: 1MB/event. 600,000,000 events/sec.
  = 1.9×10^{22} bytes/year = **19 ZettaBytes/year**
- Downsampled: 25GB/sec = 7.88×10^{17} bytes/year = **788 PetaBytes/year**
- Downsampled further: 1050MB/sec = $3.3*10^{16}$/year = **33 PetaBytes/year**

https://home.cern/about/computing/processing-what-record

Forces Driving Data Growth

- Ubiquitous sensors and reporting:
  - Cameras, mobile computing, social media, …
- Large collaborative science projects
- Philosophy: *More Data → More Value*?

http://hyperboleandahalf.blogspot.com

Enabling Technology

- **Cheap, Scalable** Data Management Systems

Why-R3: The Core of Computing

- Techniques you learn in this class underlie many topics in computing!

Essential Queries

- **Why** take this class?
- **What** is this class all about?
- **Who** is running this?
- **How** will this class work?

What is this class all about?

- Databases?
  - What is a database?
- Database Management Systems?

Looks Like?

Platters on a Disk Drive

Why the Symbol?

Looks Like?

1956: IBM
MODEL 350
RAMAC
First Commercial
Disk Drive
5MB @ 1 ton

*"...We must immediately...**attack accounting problems** under the philosophy of handling each business **transaction as it occurs**, rather than under the present condition of **batching techniques**...."*

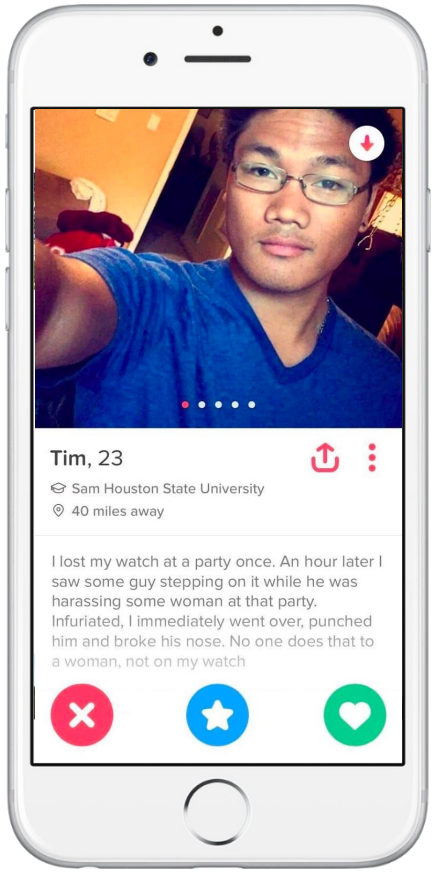
-- F. J. Wesley IBM Senior Manager

http://www.computerhistory.org/storageengine/first-commercial-hard-disk-drive-shipped

Is This a Database?

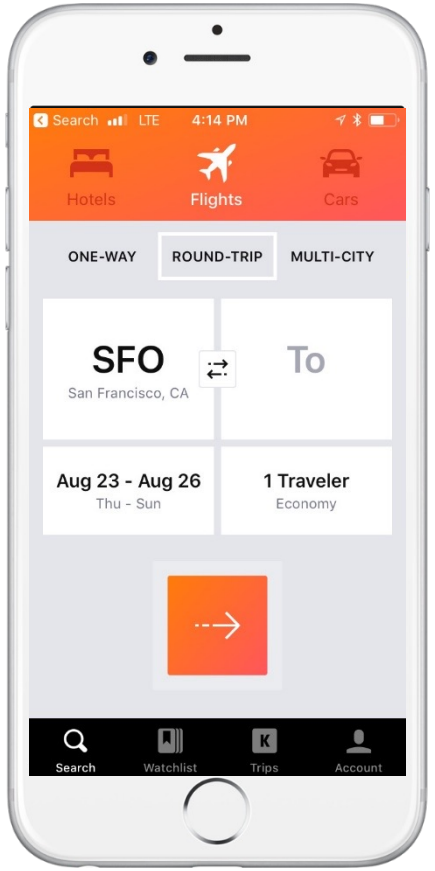
- Rolodex
- Alphbetically ordered cards
- Indexed access by first letter

Is This a Database?



- A database + “business logic” + user interface?
- Most of Tinder’s value is the database itself.

Is This a Database?

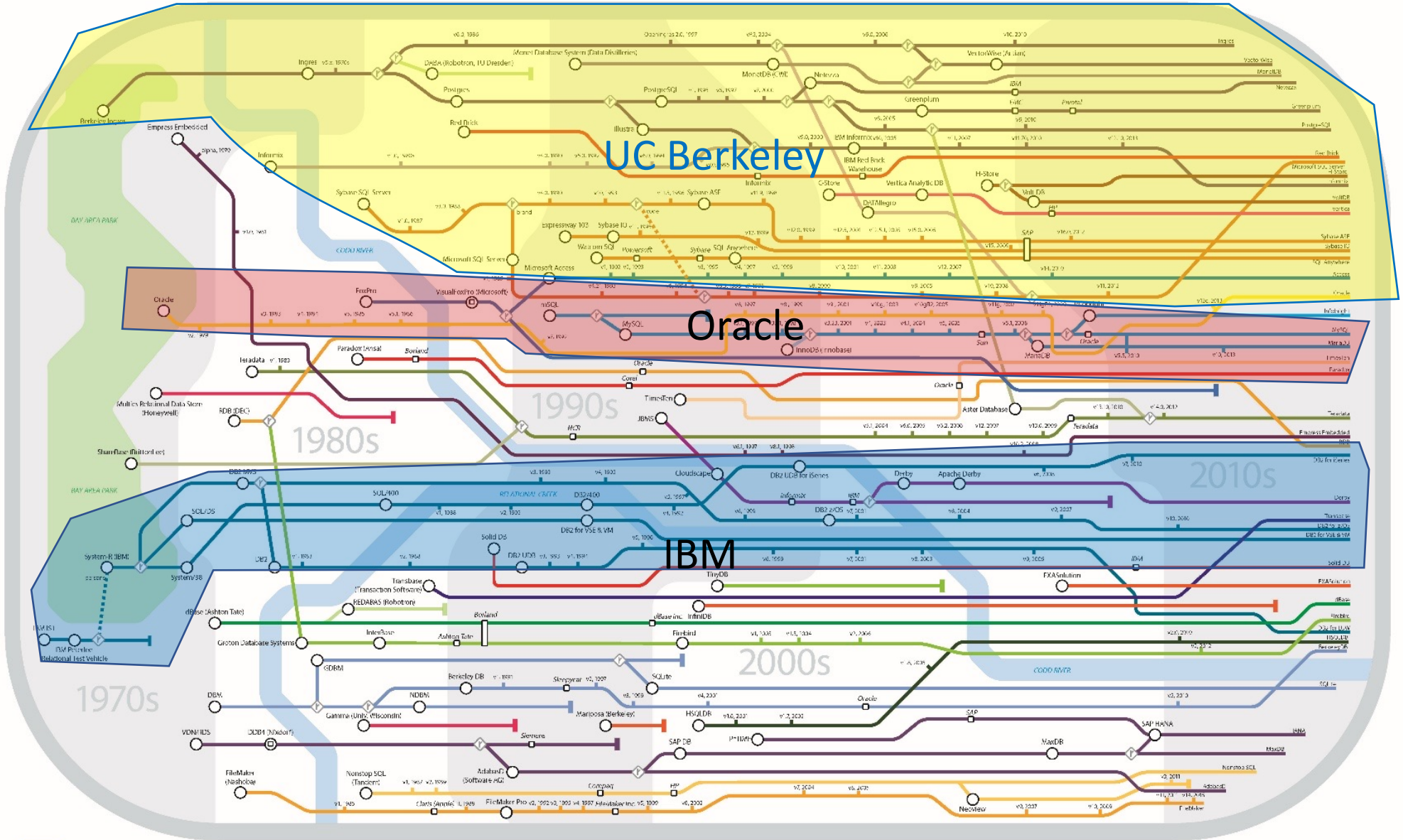


- Airline reservation systems were one of the earliest pervasive consumer uses of database systems.
  - IBM/American Airlines' SABRE system, 1964.
    - "Semi-Automated Business Research Environment"
  - Travelocity.com a direct descendant of SABRE
    - Acquired by Expedia, 1/2015

What is a Database?

- A database is a large, organized collection of data.
- Sometimes confused with a Database Management System (DBMS)
  - *A DBMS is software that **stores, manages,** and facilitates **access** to data.*

Genealogy of Relational Database Management Systems



Key to lines and symbols

Publishing Date | Acquisition | Versions | Discontinued | Branch (intellectual and/or code) | Crossing lines have no special semantics

Felix Naumann, Jana Bauckmann, Claudia Exeler, Jan-Peer Rudolph, Fabian Tschirschnitz
Contact: Hasso Plattner Institut, Potsdam, Felix.Naumann@hpi.uni-potsdam.de
Design: Alexander Sandt Grafik-Design, Hamburg
Version 5.0 - October 2015
http://www.hpi.uni-potsdam.de/naumann/projekte/rdbms_genealogy.html

Relational DBMSs

- Traditionally DBMS referred to relational databases

- **RDBMS** is a more appropriate term
- **SQL** data description and manipulation language
- **ACID** transaction consistency
- **Durable** writes (prevent data loss)
- **Mature** technologies …

Ranking of DBMS Technologies 2023

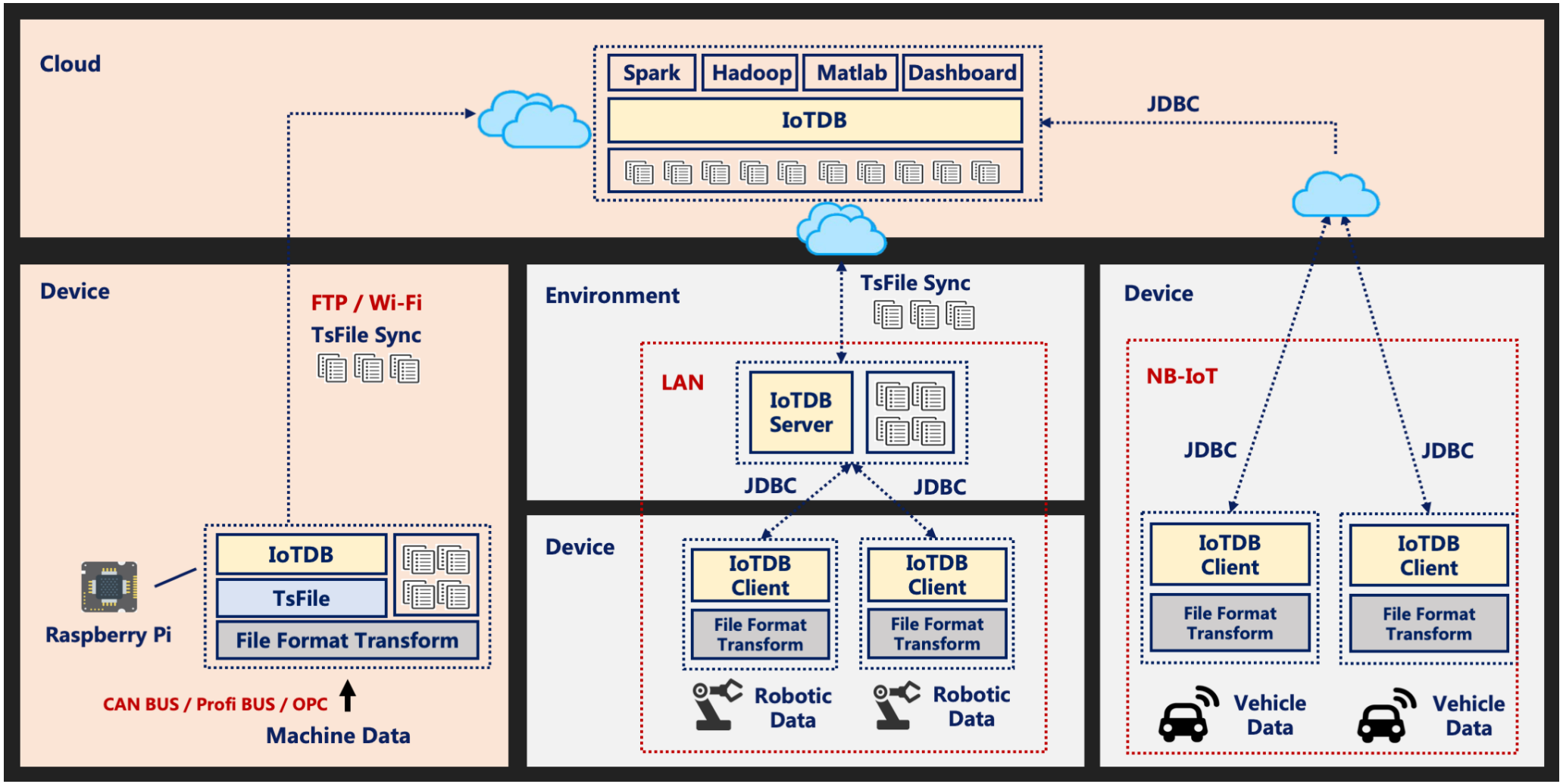
Rank			DBMS	Database Model	Score		
Sep 2023	Aug 2023	Sep 2022			Sep 2023	Aug 2023	Sep 2022
1.	1.	1.	Oracle	Relational, Multi-model	1240.88	-1.22	+2.62
2.	2.	2.	MySQL	Relational, Multi-model	1111.49	-18.97	-100.98
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	902.22	-18.60	-24.08
4.	4.	4.	PostgreSQL	Relational, Multi-model	620.75	+0.37	+0.29
5.	5.	5.	MongoDB	Document, Multi-model	439.42	+4.93	-50.21
6.	6.	6.	Redis	Key-value, Multi-model	163.68	+0.72	-17.79
7.	7.	7.	Elasticsearch	Search engine, Multi-model	138.98	-0.94	-12.46
8.	8.	8.	IBM Db2	Relational, Multi-model	136.72	-2.52	-14.67
9.	↑ 10.	↑ 10.	SQLite	Relational	129.20	-0.72	-9.62
10.	↓ 9.	↓ 9.	Microsoft Access	Relational	128.56	-1.78	-11.47

Change = Opportunity !

- The DBMS world is rapidly changing
  - Will discuss these changes towards end of the course
  - Our textbook is rather out of date
- Opportunity!
  - You can shape the future of DBMSs

- We won’t follow the textbook slavishly.

Example: IoTDB-THU



Focus & Goal

- Focus: **Foundational System Principles**
  - Basic ideas and components
  - How to compose those components into a technology stack

- Goal:
  - You will be able to **use** existing & **build new** DBMS technologies!

Databases

Main Contents

In this course, we will learn the **basic concepts**, **principles** and **applications** of database systems, especially the ***relational*** database systems. The contents mainly include :

- The data models (Ch2)
- SQL language and user interfaces (Ch3)
- Key principles of DBMS (mainly architecture, query optimization, concurrency control, recovery, etc.) (Ch4)
- The security and integrity constrains of database (Ch5)
- Database Design (Ch6)
- New Research and Application Fields

Table of Contents

1. **Introduction**

   The history, classification, and main research contents of database systems; The database system; the concepts of data model

2. **Data Model***

   Hierarchical and network model; Relational model; ER model; Object-Oriented model and other data models

3. **User Interfaces and SQL Language***

   User interface; SQL language, including QL, DDL, DCL, DML, view, embedded SQL and dynamic SQL, etc.

Table of Contents

4. **Database Management Systems***

   The architecture of database systems, query optimization, file structure and index, transaction management, concurrency control, recovery mechanism

5. **The Security and Integrity Constrain**

   The security model of database system; Integrity constrain and its expression, implementing method, assertion, trigger

6. **Database Design***

   Design procedure; ER graph; Normalization of Relational Schema

1. Introduction

What Is Database?

What Is Database management System (DBMS)?

Database & DBMS

- A very large, integrated collection of data.

- Models real-world *enterprise.*
  - Entities (e.g., students, courses)
  - Relationships (e.g., electives)

- A *Database Management System (DBMS)* is a software package designed to store and manage databases.

Any experience of dealing with data?

Files?

Files vs. Databases

- Application must stage large datasets between main memory and secondary storage (e.g., buffering, page-oriented access, 32-bit addressing, etc.)
- Special code for different queries
- Must protect data from inconsistency due to multiple concurrent users
- Crash recovery
- Security and access control

Purpose of Database Systems

In the early days, database applications were built directly on top of file systems, which leads to:

- Data redundancy and inconsistency: data is stored in multiple file formats resulting induplication of information in different files
- Difficulty in accessing data
  - Need to write a new program to carry out each new task
- Data isolation
  - Multiple files and formats
- Integrity problems
  - Integrity constraints (e.g., account balance > 0) become "buried" in program code rather than being stated explicitly
  - Hard to add new constraints or change existing ones

Why Use a DBMS?

Why Use a DBMS?

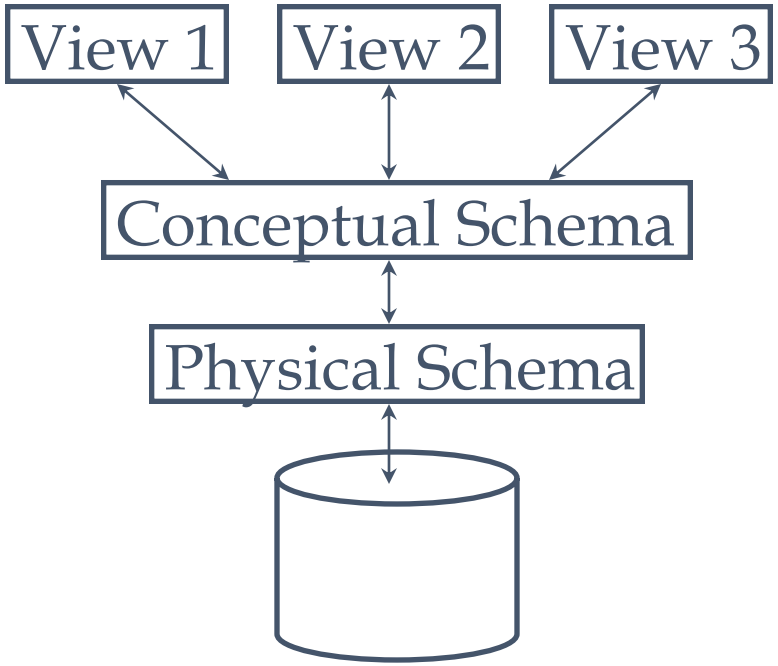
- Data independence and efficient access.
- Reduced application development time.
- Data integrity and security.
- Uniform data administration.
- Concurrent access, recovery from crashes.

Data, Data Model and Data Schema

- *Data* are symbols for describing the things of real world. They are existing form of information.
- A *data model* is a collection of concepts and definitions for describing data.
- A *schema* is a description of a particular collection of data, using a given data model.
- The *relational model of data* is the most widely used model today.
  - Main concept: *relation*, basically a table with rows and columns.
  - Every relation has a *schema*, which describes the columns, or fields.

Levels of Abstraction: ANSI-SPARC Architecture

- Many *views*, single *conceptual (logical) schema* and *physical schema*.
  - Views describe how users see the data.
  - Conceptual schema defines logical structure
  - Physical schema describes the files and indexes used.



☛ *Schemas are defined using DDL; data is modified/queried using DML.*

Data Definition Language (DDL)

- Specification notation for defining the database schema

  Example: **create table** *instructor* (
  ID **char**(5),
  name **varchar**(20)**,**
  dept_name **varchar**(20),
  salary **numeric**(8,2))

- DDL compiler generates a set of table templates stored in a ***data dictionary***
- Data dictionary contains metadata (i.e., data about data)
  - Database schema
  - Integrity constraints
    - Primary key (ID uniquely identifies instructors)
  - Authorization
    - Who can access what

Data Manipulation Language (DML)

- Language for accessing and updating the data organized by the appropriate data model
  - DML also known as **query language**
- There are basically two types of data-manipulation language
  - **Procedural DML** -- require a user to specify what data are needed and how to get those data.
  - **Declarative DML** -- require a user to specify what data are needed without specifying how to get those data.
- Declarative DMLs are usually easier to learn and use than are procedural DMLs.
- Declarative DMLs are also referred to as non-procedural DMLs
- The portion of a DML that involves information retrieval is called a **query** language.

SQL Query Language

- SQL query language is nonprocedural. A query takes as input several tables (possibly only one) and always returns a single table.
- Example to find all instructors in Comp. Sci. dept

```
select name
from instructor
where dept_name = 'Comp. Sci.'
```

- SQL is **NOT** a Turing machine equivalent language
  - To be able to compute complex functions SQL is usually embedded in some higher-level language
- Application programs generally access databases through one of
  - Language extensions to allow embedded SQL
  - Application program interface (e.g., ODBC/JDBC) which allow SQL queries to be sent to a database

Database Access from Application Program

- Non-procedural query languages such as SQL are not as powerful as a universal Turing machine.
- SQL does not support actions such as input from users, output to displays, or communication over the network.
- Such computations and actions must be written in a **host language**, such as C/C++, Java or Python, with embedded SQL queries that access the data in the database.
- **Application programs** -- are programs that are used to interact with the database in this fashion.

Example: University Database

- Conceptual schema:
  - *Students(sid: string, name: string, login: string, age: integer, gpa:real)*
  - *Courses(cid: string, cname:string, credits:integer)*
  - *Enrolled(sid:string, cid:string, grade:integer)*
- Physical schema:
  - Relations stored as unordered files.
  - Index on first column of Students.
- External Schema (View):
  - *Course_info(cid:string,enrollment:integer)*