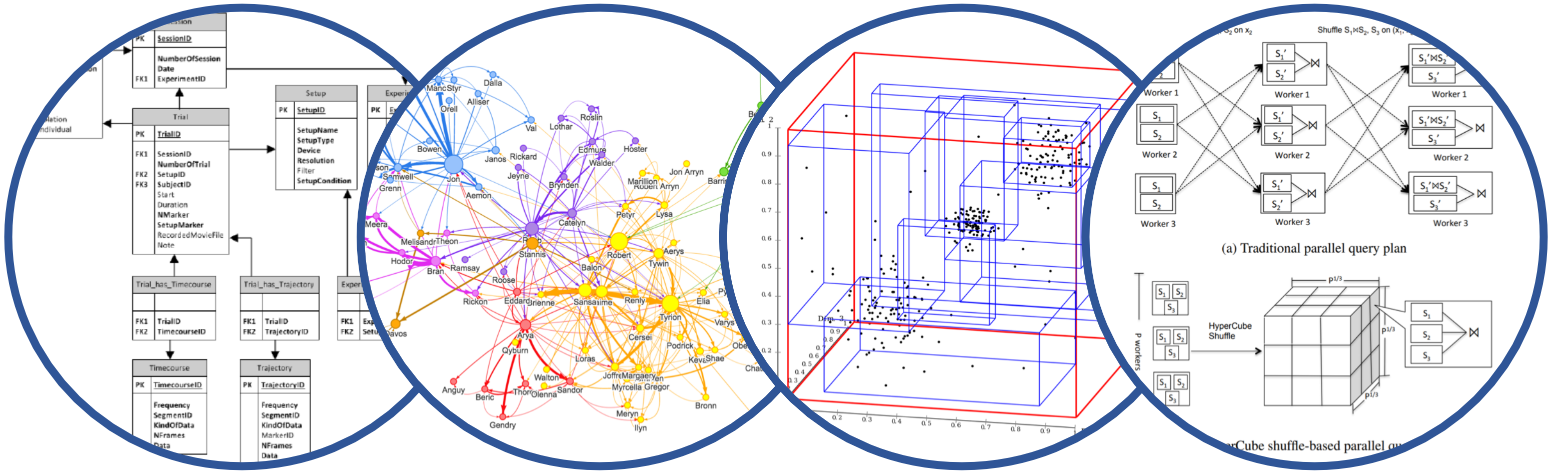


Introduction to Data Management

Practical Data Management

Paul G. Allen School of Computer Science and Engineering
University of Washington, Seattle

Announcements

- Midterm exam
  - In-class on Wednesday
  - 1 double-sided page of notes, printed if you wish
  - Covers all material up until last Wednesday (Yes E/R diagrams and functional dependencies. No BCNF decomp.)

- Practice exams (only for material we've covered):
  - Last quarter's exam and solutions
    https://sites.google.com/cs.washington.edu/cse344-2019sp/home
  - 414 exam from last autumn
    https://courses.cs.washington.edu/courses/cse414/18au/exams.html
  - You can use other past tests for reference but they may not represent what's on our exam. We've also found errors in previous pdfs that were never fixed!

Goals for Today

- Finish design theory content on BCNF
- Talk about the fuzzy stuff in data management
  - Data cleaning
  - Private data and ethics

Outline

- Data Cleaning
  - ETL
  - Data wrangling on GCP Dataprep (Trifacta)
- Data Management Ethics and Best Practices

Where is my data coming from?

Mainly two possible sources:

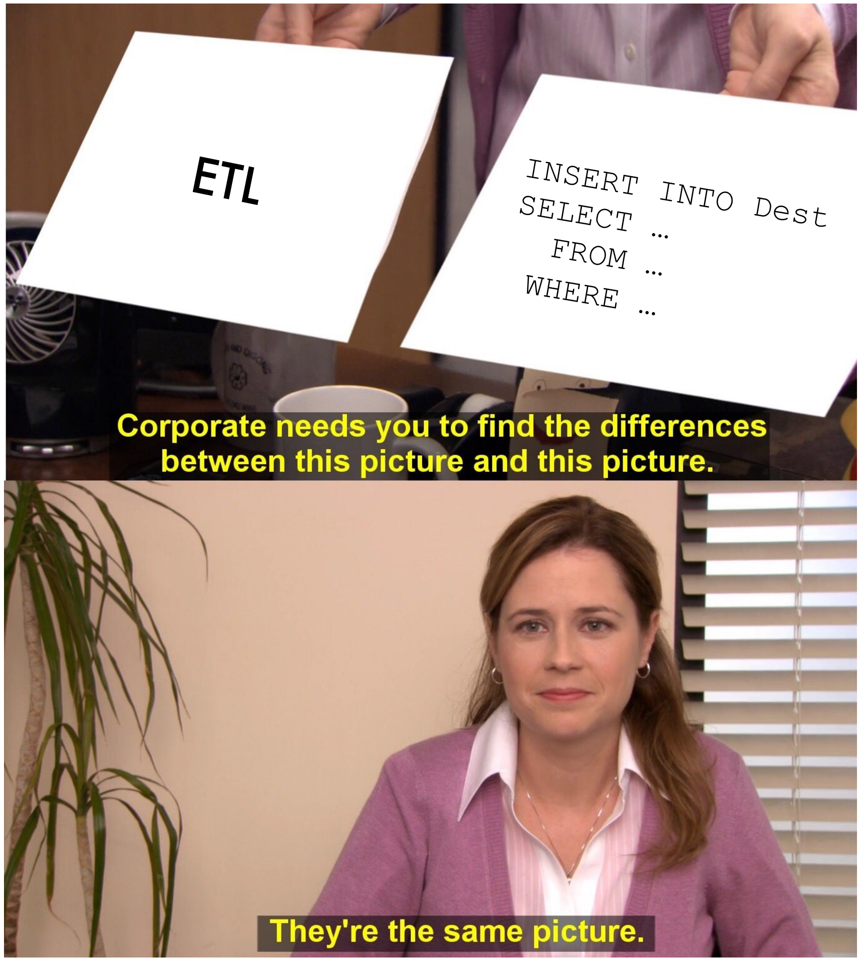
- You generate the data
  - Output data that is easy to use
- External sources or preexisting data
  - Sometimes doesn’t fit your application needs
  - Need to translate the data into a usable form

Extract Transform Load (ETL)

“I know exactly what operations need to be done to get from data format A to data format B”

- Extract
  - Read relevant data
- Transform
  - Push data through mapping functions until done
    - Aggregations
    - Normalization
    - …
- Load
  - Write to destination

Extract Transform Load (ETL)


ETL
INSERT INTO Dest
SELECT …
FROM …
WHERE …
Corporate needs you to find the differences between this picture and this picture.
They're the same picture.

Data Wrangling

“I have no clue what’s going on with my data“

- Essentially ETL but with **data exploration**
- Interactivity is important
  - Visualizations
  - Suggestions

Pivot

- Create a "summary table"
  - Generally used for reports to draw attention to interesting values
  - Able to make values into columns
- "Skinny and tall" → "short and wide"

Name	Year	GDP
Angola	2015	100
Luxembourg	2015	50
Angola	2016	110
Angola	2018	115
Luxembourg	2017	55
Luxembourg	2018	65

Pivot

- Create a "summary table"
  - Generally used for reports to draw attention to interesting values
  - Able to make values into columns
- "Skinny and tall" → "short and wide"

GDP relation:

Name	2015	2016	2017	2018
Angola	100	110		115
Luxembourg	50		55	65

Unpivot

- Usually we want to store unpivoted data
  - Easier to manage
- “Short and wide” → “skinny and tall”

GDP relation:

Name	2015	2016	2017	2018
Angola	100	110		115
Luxembourg	50		55	65

Data Wrangling

Cloud Dataprep by TRIFACTA

TRIFACTA

google-refine → OpenRefine

TIBCO™ Clarity

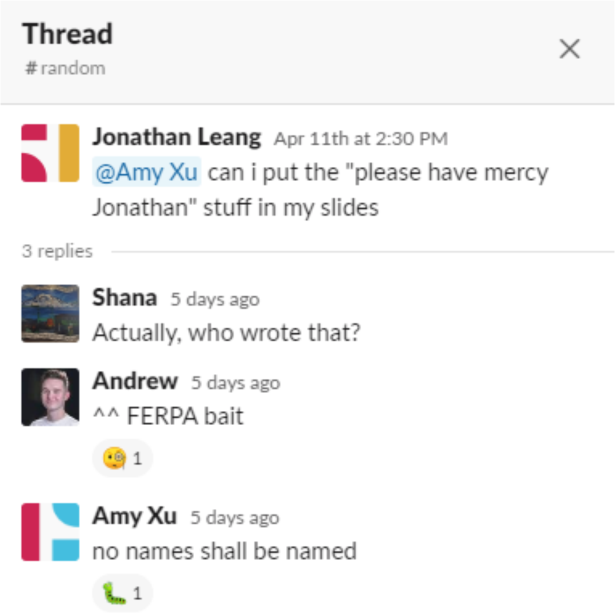
alteryx

Now what?

You can get data but what are you doing with it?

Existing Laws and Regulations

- FERPA (Family Education Rights and Privacy Act)
- Mandatory for education institutions
  - Requires written consent to disclose academic info
  - Allows the release of directory information



Jonathan Leang Apr 11th at 2:30 PM
@Amy Xu can i put the "please have mercy Jonathan" stuff in my slides

3 replies

Shana 5 days ago
Actually, who wrote that?

Andrew 5 days ago
^^ FERPA bait
🧐 1

Amy Xu 5 days ago
no names shall be named
🐛 1

Existing Laws and Regulations

- HIPAA (Health Information Portability and Accountability Act)
- Mandatory for healthcare and health insurance institutions
  - Privacy Rule to protect Protected Health Information
  - Security Rule to ensure administrative, physical, and technical safeguards

Existing Laws and Regulations

- GDPR (General Data Protection Regulation)
- Recently became law in the EU
  - Requires disclosure by companies on how they use user information
  - Last year all of the US tried to become compliant.. almost everyone waited until the deadline

- Extremely important to follow these protocols!

Sensitive Information

- Personal identifiers
  - Names
  - Student ID
  - Social security number
  - License number
- Protected data (for legal and/or ethical reasons)
  - Academic records (FERPA)
  - Protected Health Information (HIPAA)
- Passwords

Access Control

- Block people who shouldn't have access
  - Most large companies have a tiered-access hierarchy
- Databases usually have built-in access control:

```
GRANT <permissions>
  [ON <table>]
   TO <user/role>
```

```
GRANT SELECT, INSERT
   ON MySecureTable
   TO PUBLIC
```

Allow anyone who can connect to read and add data to MySecureTable

Permissions:

- Table-level operations (SELECT, DELETE, …)
- DB-level operations (CREATE TABLE, GRANT, …)

User/Role:

- Users like a user on your computer
- Roles (groups) can be predefined or created

Access Control

- SQL Injection → application input acts as code
  - Union attack, tautology attack, illegal queries
  - Only possible if there is a place to inject code
  - Consistently one of the top web-based attacks
    - People simply don't realize its an issue or…
    - People know it's an issue and never get around to fixing it
- Considered a "solved" problem
  - **Parameterize queries with prepared statements**

Access Control

Other common techniques to limit access:

- Limit the number of rows that can be seen
  - Leaking a few tuples is better than leaking all of them
- Only allow aggregations
  - Grouping implicitly eliminates identification info
- Don’t store data you don’t need!

Anonymize Data

FERPA Deidentification

- ID to anonymous ID mapping should be secret
- Aggregate data (minimum n-size)
  - **Suppression** → Don’t provide data ☹
    - Necessary for very small groups
  - **Rounding** → Bucket data or introduce noise ☺
    - More people means you can be more specific

Implicit Disclosure

- FERPA allows institutions to disclose “directory information” without consent (institution policies can be stronger)
  - Name
  - Email
  - Photographs
  - Phone Number
- If users can derive sensitive information like grades, it violates FERPA

Implicit Disclosure

- "Hey, can you give me the directory information for students with a GPA of 3.5?"

Implicit Disclosure

- "Hey, can you give me the directory information for students with a GPA of 3.5?"

Reveals sensitive information by context

```
SELECT D.*
  FROM Directory AS D, Grades AS G
 WHERE D.id = G.id AND
       G.gpa = 3.5
```

Implicit Disclosure

Re-identification of Mass. Governor William Weld

- Public voter data
  - Name
  - ZIP code
  - Sex
  - Birth date
  - …
- Anonymous insurance data
  - ZIP code
  - Sex
  - Birth date
  - Prescription
  - Diagnosis
  - …

Implicit Disclosure

Cambridge, MA Voter Data ($20)

Name	ZIP	Sex	Bday
…	…	…	…
W. Weld	12345	M	Feb 30
…	…	…	…

$\bowtie$

Anon. Insurance Data for Researchers

ZIP	Sex	Bday	MedInfo
…	…	…	…
12345	M	Feb 30	Affluenza
…	…	…	…

6 matches on ZIP
3 matches on Sex
1 match on Bday

Name	…	MedInfo
…	…	…
W. Weld	…	Affluenza
…	…	…

Implicit Disclosure

Cambridge, MA Voter Data ($20)

Name	ZIP	Sex	Bday
…	…	…	…
W. Weld	12345	M	Feb 30
…	…	…	…

$\bowtie$

Anon. Insurance Data for Researchers

ZIP	Sex	Bday	MedInfo
…	…	…	…
12345	M	Feb 30	Affluenza
…	…	…	…

6 matches on ZIP
3 matches on Sex
1 match on Bday

Name	…	MedInfo
…	…	…
W. Weld	…	Affluenza
…	…	…

Implicit Disclosure

Cambridge, MA Voter Data ($20)

Name	ZIP	Sex	Bday
…	…	…	…
W. Weld	12345	M	Feb 30
…	…	…	…

⋈

Anon. Insurance Data for Researchers

ZIP	Sex	Bday	MedInfo
…	…	…	…
12345	M	Feb 30	Afluenza
…			

Legal in 1997
Illegal since 2003

6 matches on ZIP
3 matches on Sex
1 match on Bday

Name	…	MedInfo
…	…	…
W. Weld	…	Afluenza
…	…	…

Storing Passwords

- Passwords are special
  - High potential for additional security compromises
  - Only operation that should be done is equality comparison

Storing Passwords

(bobtheninja246, password)

If you do this, Ted Codd will start rolling in his grave.

Username	Password
bobtheninja246	password
xXxDragonSlayerxXx	password
420_E-Sports_Masta	qwertyuiop

Storing Passwords

- Quick overview of hashing
  - Hash(input) → hash value
  - Hashing is <u>deterministic</u>
  - Ideally hashing is <u>noninverible</u>
  - Ideally hash values are uniformly spread out

Storing Passwords

Hash it!

(bobtheninja246, hash(password))

(bobtheninja246, FCgJFI9ryz)

Username	Hash
bobtheninja246	FCgJFI9ryz
xXxDragonSlayerxXx	FCgJFI9ryz
420_E-Sports_Masta	p8mel6usIF

Storing Passwords

Hash it!

(bobtheninja246, hash(password))

(bobtheninja246, FCgJFI9ryz)

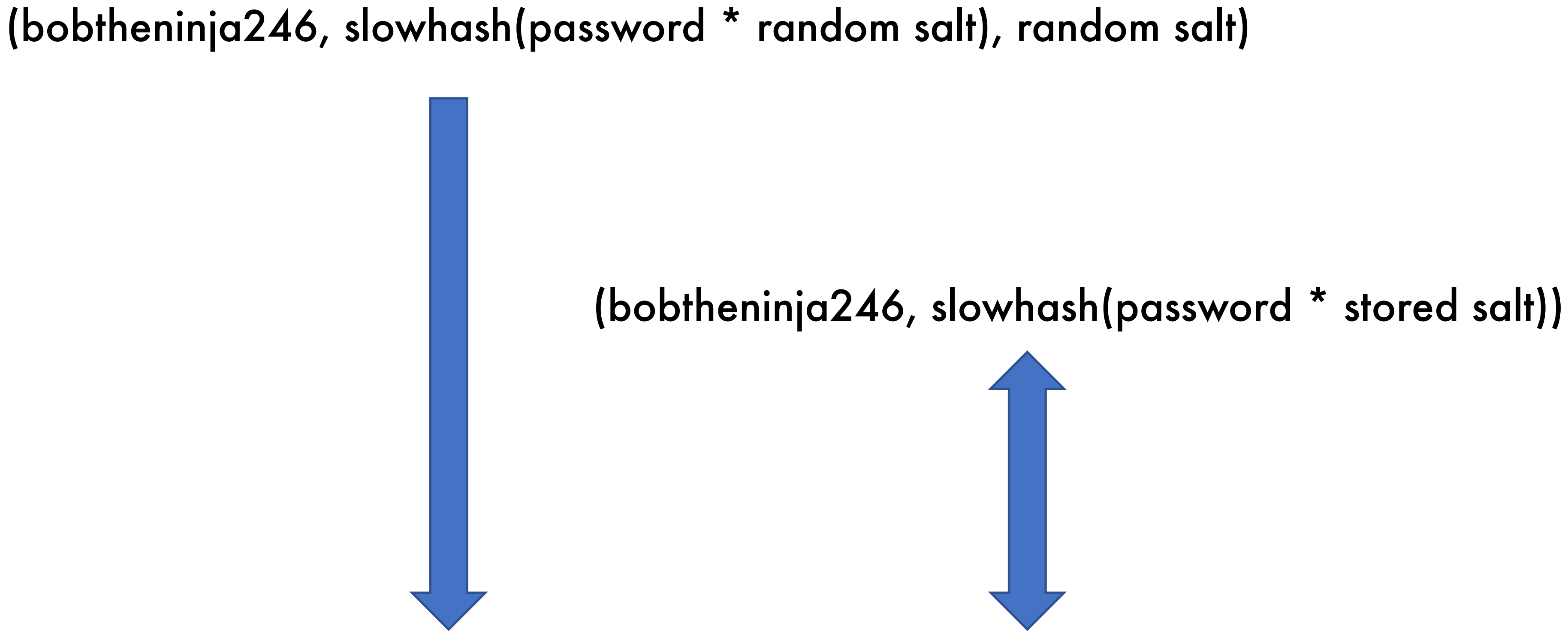
Issues/pitfalls:

- Hashing functions have precomputed "rainbow tables"
- Some hashing functions are fast so brute forcing attacks can happen
- Patterns can occur for the same passwords

Username	Hash
bobtheninja246	FCgJFI9ryz
xXxDragonSlayerxXx	FCgJFI9ryz
420_E-Sports_Masta	p8mel6usIF

Storing Passwords

Salt it and hash it!



(bobtheninja246, slowhash(password * random salt), random salt)

(bobtheninja246, slowhash(password * stored salt))

Username	Hash	Salt
bobtheninja246	HHxrd5o7Cn	WUKhhIFBLc
xXxDragonSlayerxXx	7rYFQIowpW	mq5rFL6JzF
420_E-Sports_Masta	cQF4DdSFfn	S8e0zpATNR

Storing Passwords

Salt it and hash it!

(bobtheninja246, slowhash(password * random salt), random salt)

stored salt))

These are just the fundamentals!
Many companies outsource password management because it can get very complicated.
In real applications never roll your own protocol!

Username	Hash	Salt
bobtheninja246	HHxrd5o7Cn	WUKhhIFBLc
xXxDragonSlayerxXx	7rYFQlowpW	mq5rFL6JzF
420_E-Sports_Masta	cQF4DdSFfn	S8e0zpATNR

Data Quality

- Quality is not only about cleanness
- Quality may also involve significance
  - Are certain groups large enough to draw meaningful aggregates?
  - If my data is a sample of a population, does it accurately depict that population?

Worlds Shortest Intro to Machine Learning

- Training data → Prediction program
  - Prediction program believes that the training data is representative of a population and covers all cases