Introduction to Data Management

Practical Data Management

Alyssa Pittman
Based on slides by Jonathan Leang, Dan Suciu, et al

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

- Midterm a week from today!
  - Allowed 1 handwritten sheet of notes (both sides)
  - Content that is fair game:
    - any content through today’s lecture
    - lecture + readings + hw
    - no questions on Datalog aggregates/groups

- Monday is midterm review in-class
  - Bring questions! Re content or past midterm q’s

- My office hours cancelled next week
  - Schedule 1-1 if you need me or see TA OHs
Goals for Today

- Talk about the parts of data management you might encounter in the real world without having theory to back you up
Outline

▪ Data Cleaning
  • ETL
    • Data wrangling on GCP Dataprep (Trifacta)
▪ Data Management Ethics and Best Practices
Where is my data coming from?

- 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
“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)

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”

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angola</td>
<td>2015</td>
<td>100</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>2015</td>
<td>50</td>
</tr>
<tr>
<td>Angola</td>
<td>2016</td>
<td>110</td>
</tr>
<tr>
<td>Angola</td>
<td>2018</td>
<td>115</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>2017</td>
<td>55</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>2018</td>
<td>65</td>
</tr>
</tbody>
</table>
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”

<table>
<thead>
<tr>
<th>Name</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angola</td>
<td>100</td>
<td>110</td>
<td></td>
<td>115</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>50</td>
<td></td>
<td>55</td>
<td>65</td>
</tr>
</tbody>
</table>
Unpivot

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

<table>
<thead>
<tr>
<th>Name</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angola</td>
<td>100</td>
<td>110</td>
<td></td>
<td>115</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>50</td>
<td></td>
<td>55</td>
<td>65</td>
</tr>
</tbody>
</table>
Data Wrangling

Quickstart - demo

Cloud Dataprep by 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
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

- SOX (Sarbanes–Oxley Act)
- Requires auditability for companies’ financial records
  - What does this have to do with data?
  - Can financial data be tampered with?
  - Can code touching financial data be tampered with?
Existing Laws and Regulations

- **GDPR (European General Data Protection Regulation)**
  - Requires disclosure from companies about what user data they have and how they use it
    - ...but can be *exploited*
Laws and Regulations Today

- Social Media and Politics?
- Facebook–Cambridge Analytica Scandal
  - CA uses loophole in Facebook API through an online quiz to harvest personal information data

Whistleblower Christopher Wylie

Mark Zuckerberg’s hearing
What’s at Stake?

Jane Lytvynenko
@JaneLytv

The details from his Equifax class-action suit are BONKERS

That these weak passwords had already been compromised in previous breaches. Furthermore, Equifax employed the username “admin” and the password “admin” to protect a portal used to manage credit disputes, a password that “is a surefire way to get hacked.” This portal contained a vast trove of personal information. According to cybersecurity experts, these shortcomings

9:40 AM - 18 Oct 2019
Sensitive Information

- **PII** = Personally identifying information
  - Names
  - Student ID
  - Social security number
  - License number

- Protected data (for legal and/or ethical reasons)
  - Academic records (FERPA)
  - Protected Health Information (HIPAA)
  - Customer records (GLBA)

- 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)**

<table>
<thead>
<tr>
<th>Name</th>
<th>ZIP</th>
<th>Sex</th>
<th>Bday</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>W. Weld</td>
<td>12345</td>
<td>M</td>
<td>Feb 30</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

**Anon. Insurance Data for Researchers**

<table>
<thead>
<tr>
<th>ZIP</th>
<th>Sex</th>
<th>Bday</th>
<th>MedInfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>12345</td>
<td>M</td>
<td>Feb 30</td>
<td>Afluenza</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

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

**Name** | **…** | **MedInfo**
---|---|---
… | … | …
W. Weld | … | Afluenza
… | … | …
Implicit Disclosure

Cambridge, MA Voter Data ($20)

<table>
<thead>
<tr>
<th>Name</th>
<th>ZIP</th>
<th>Sex</th>
<th>Bday</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>W. Weld</td>
<td>12345</td>
<td>M</td>
<td>Feb 30</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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

Anon. Insurance Data for Researchers

<table>
<thead>
<tr>
<th>ZIP</th>
<th>Sex</th>
<th>Bday</th>
<th>MedInfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>12345</td>
<td>M</td>
<td>Feb 30</td>
<td>Afluenza</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Legal in 1997
Illegal since 2003
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.

<table>
<thead>
<tr>
<th>Username</th>
<th>Password</th>
</tr>
</thead>
<tbody>
<tr>
<td>bobtheninja246</td>
<td>password</td>
</tr>
<tr>
<td>xXxDragonSlayerxXx</td>
<td>password</td>
</tr>
<tr>
<td>420_E-Sports_Masta</td>
<td>qwertyuiop</td>
</tr>
</tbody>
</table>
Quick overview of hashing
• Hash(input) → hash value
• Hashing is deterministic
• Ideally hashing is noninvertible
• Ideally hash values are uniformly spread out
Storing Passwords

Hash it!

\[(\text{bobtheninja246}, \text{hash(password)})\]

\[(\text{bobtheninja246}, \text{FCgJFI9ryz})\]

<table>
<thead>
<tr>
<th>Username</th>
<th>Hash</th>
</tr>
</thead>
<tbody>
<tr>
<td>bobtheninja246</td>
<td>FCgJFI9ryz</td>
</tr>
<tr>
<td>xXxDragonSlayerxXx</td>
<td>FCgJFI9ryz</td>
</tr>
<tr>
<td>420_E-Sports_Masta</td>
<td>p8mel6usIF</td>
</tr>
</tbody>
</table>
Hash it!

\[(\text{bobtheninja246}, \text{hash(password)})\]

\[(\text{bobtheninja246}, \text{FCgJFI9ryz})\]

**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

<table>
<thead>
<tr>
<th>Username</th>
<th>Hash</th>
</tr>
</thead>
<tbody>
<tr>
<td>bobtheninja246</td>
<td>FCgJFI9ryz</td>
</tr>
<tr>
<td>xXxDragonSlayerxXx</td>
<td>FCgJFI9ryz</td>
</tr>
<tr>
<td>420_E-Sports_Masta</td>
<td>p8mel6usIF</td>
</tr>
</tbody>
</table>
Storing Passwords

Salt it and hash it!

\[(\text{bobtheninja246, slowhash(password} \times \text{random salt}), \text{random salt})\]

<table>
<thead>
<tr>
<th>Username</th>
<th>Hash</th>
<th>Salt</th>
</tr>
</thead>
<tbody>
<tr>
<td>bobtheninja246</td>
<td>HHxrd5o7Cn</td>
<td>WUKhhIFBLc</td>
</tr>
<tr>
<td>xXxDragonSlayerxXx</td>
<td>7rYFQIowpW</td>
<td>mq5rFL6JzF</td>
</tr>
<tr>
<td>420_E-Sports_Masta</td>
<td>cQF4DdSFfn</td>
<td>S8e0zpATNR</td>
</tr>
</tbody>
</table>
Storing Passwords

Salt it and hash it!

$$\text{(bobtheninja246, slowhash(password * random salt), random salt)}$$

These are just the fundamentals! Many companies outsource password management because it can get very complicated.

<table>
<thead>
<tr>
<th>Username</th>
<th>Hash</th>
<th>Salt</th>
</tr>
</thead>
<tbody>
<tr>
<td>bobtheninja246</td>
<td>HHxrd5o7Cn</td>
<td>WUKhhIFBLc</td>
</tr>
<tr>
<td>xXxDragonSlayerxXx</td>
<td>7rYFQIowpW</td>
<td>mq5rFL6JzF</td>
</tr>
<tr>
<td>420_E-Sports_Masta</td>
<td>cQF4DdSFfn</td>
<td>S8e0zpATNR</td>
</tr>
</tbody>
</table>
Data Quality

- Quality is not only about cleanliness
- 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?
  - Did I ask the right kinds of questions to get good data?
Even Affects Machine Learning

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

- Be good stewards of the data you have
- There's more to data management than the technical bits