CS 327E Intro

January 28, 2019
Terminology

• Dataset
• Relation / Entity Type / Table
• Field / Attribute / Column
• Row / Entity / Tuple / Record
• Cell / Value
• Data Type (e.g. INT, NUMERIC, STRING, BOOL, DATE, TIMESTAMP)
• Constraint (e.g. NOT NULL, Primary Key, Foreign Key)
• Schema
• Database
Table Relationship: One-to-Many (1:m)
Table Relationship: One-to-Many (1:m)

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>section</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mary Tuma</td>
<td>news</td>
</tr>
<tr>
<td>2</td>
<td>Michael King</td>
<td>arts</td>
</tr>
<tr>
<td>3</td>
<td>Nina Hernandez</td>
<td>news</td>
</tr>
<tr>
<td>4</td>
<td>Sunil Kumar</td>
<td>music</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>date</th>
<th>authid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Turmoil at the Zoo</td>
<td>2019-01-26</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>CodeNEXT’s New Friend</td>
<td>2019-01-27</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Quote of the Week</td>
<td>2019-01-27</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>SXSW News</td>
<td>2019-01-28</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>More from Steve Adler</td>
<td>2019-01-28</td>
<td>1</td>
</tr>
</tbody>
</table>
Table Relationship: One-to-One (1:1)
Table Relationship: One-to-One (1:1)

![Diagram showing a one-to-one relationship between two tables, T1 and T2, with fields field1, field2, field3, and fieldn.]

**Article**

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>date</th>
<th>authid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Turmoil at the Zoo</td>
<td>2019-01-26</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>CodeNEXT's New Friend</td>
<td>2019-01-27</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Quote of the Week</td>
<td>2019-01-27</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>SXSW News</td>
<td>2019-01-28</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>More from Steve Adler</td>
<td>2019-01-28</td>
<td>1</td>
</tr>
</tbody>
</table>

**Article_Stats**

<table>
<thead>
<tr>
<th>id</th>
<th>clicks</th>
<th>likes</th>
<th>dislikes</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120</td>
<td>45</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Table Relationship: Many-to-Many ($m:n$)
Table Relationship: Many-to-Many (m:n)

Tag

<table>
<thead>
<tr>
<th>id</th>
<th>tag</th>
<th>aids</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Politics</td>
<td>1, 2, 5</td>
</tr>
<tr>
<td>2</td>
<td>Austin</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>3</td>
<td>Mayor</td>
<td>3, 5</td>
</tr>
<tr>
<td>4</td>
<td>Business</td>
<td>1, 2, 5</td>
</tr>
<tr>
<td>6</td>
<td>Land Development</td>
<td>2</td>
</tr>
<tr>
<td>37</td>
<td>Animals</td>
<td>1</td>
</tr>
</tbody>
</table>

Article

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>date</th>
<th>authid</th>
<th>tids</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Turmoil at the Zoo</td>
<td>2019-01-26</td>
<td>1</td>
<td>4, 37, 2</td>
</tr>
<tr>
<td>2</td>
<td>CodeNEXT’s New Friend</td>
<td>2019-01-27</td>
<td>1</td>
<td>2, 6</td>
</tr>
<tr>
<td>3</td>
<td>Quote of the Week</td>
<td>2019-01-27</td>
<td>3</td>
<td>2, 3</td>
</tr>
<tr>
<td>4</td>
<td>SXSW News</td>
<td>2019-01-28</td>
<td>2</td>
<td>2, 40, 7</td>
</tr>
<tr>
<td>5</td>
<td>More from Steve Adler</td>
<td>2019-01-28</td>
<td>1</td>
<td>2, 3, 4</td>
</tr>
</tbody>
</table>
Representation of Many-to-Many \((m:n)\)
# Table Relationship: Many-to-Many \((m:n)\)

## Tag

<table>
<thead>
<tr>
<th>id</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Politics</td>
</tr>
<tr>
<td>2</td>
<td>Austin</td>
</tr>
<tr>
<td>3</td>
<td>Mayor</td>
</tr>
<tr>
<td>4</td>
<td>Business</td>
</tr>
<tr>
<td>6</td>
<td>Land Development</td>
</tr>
<tr>
<td>37</td>
<td>Animals</td>
</tr>
</tbody>
</table>

## Article

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
<th>date</th>
<th>authid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Turmoil at the Zoo</td>
<td>2019-01-26</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>CodeNEXT's New Friend</td>
<td>2019-01-27</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Quote of the Week</td>
<td>2019-01-27</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>SXSW News</td>
<td>2019-01-28</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>More from Steve Adler</td>
<td>2019-01-28</td>
<td>1</td>
</tr>
</tbody>
</table>

## Tagged_Article

<table>
<thead>
<tr>
<th>tid</th>
<th>aid</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>
Have you ever spent a lot of time on a computer program, made lots of dumb mistakes, and eventually fixed things with some insight?

A. Yes, I’ve been there
B. No, not really
C. I never write buggy code
Term Project

- Create a data warehouse
- 13 Milestones
- **Milestone 1** due Friday
Datasets

• A set of related data files produced from the same source
• Dataset types: main dataset and secondary datasets
• Choose data you want to analyze to gain some insights
Main Dataset

- AKA *Dataset1*
- Comprised of $N$ files ($3 < N < 30$)
- CSV format
- At least 10K rows per file
- At least 5 columns per file
- Multiple parent/child relationships present in the data
- Dirty data
Secondary Dataset

- AKA Dataset2
- Related to your main dataset (aka Dataset1)
- Comprised of $N$ files ($3 < N < 30$)
- At least 10K rows per file
- At least 5 columns per file
- CSV format
- Dirty data
# A Few Examples

<table>
<thead>
<tr>
<th></th>
<th>Main Dataset</th>
<th>Secondary Dataset(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transportation</strong></td>
<td>Airline on-time performance (source: Bureau of Transportation Statistics)</td>
<td>Storm events (source: National Oceanic and Atmospheric Administration)</td>
</tr>
<tr>
<td><strong>Housing</strong></td>
<td>Short-term rentals in various cities (source: Airbnb)</td>
<td>Long-term rentals nationwide (source: Zillow)</td>
</tr>
<tr>
<td><strong>Employment</strong></td>
<td>H1B visa applications (source: US Dept. of Labor)</td>
<td>Corporate Registrations (source: Secretary of States)</td>
</tr>
<tr>
<td><strong>Movies</strong></td>
<td>Hollywood movies, directors, actors (source: IMDB)</td>
<td>Bollywood movies, actors and songs (source: Cinemalytics)</td>
</tr>
<tr>
<td><strong>Music</strong></td>
<td>Artists and songs (source: MusicBrainz)</td>
<td>Artists, labels, recordings on vinyl and other formats (source: Discog)</td>
</tr>
</tbody>
</table>
Main Dataset:
H1B Visa applications

Source:
US Dept. of Labor

Table Sizes:
2015 table: 241 MB size, 618,804 rows
2016 table: 233 MB size, 647,852 rows
2017 table: 253 MB size, 624,650 rows
2018 table: 283 MB size, 654,162 rows

Table Schemas:
-A few schema variations between the tables (column names, data types).

Project Work:
-Imported files into BQ tables
-Milestones 1 and 2
Project Work:
-Merged and split raw tables
-Enforced referential integrity
-Removed duplicate records
-Milestones 4, 5, 6
Secondary Dataset:
Corporate Registrations

Source:
Secretary of States

Table Sizes:
AZ: 225 MB size, 869,943 rows
CA: 1.1 GB size, 3,792,457 rows
CO: 38 MB size, 160,808 rows
CT: 192 MB size, 796,877 rows
GA: 302 MB size, 2,076,016 rows; 116 MB size, 2,063,919 rows
MA: 221 MB size, 1,066,639 rows
MN: 374 MB size, 1,688,714 rows; 799 MB size, 4,072,355 rows
MO: 133 MB size, 2,364,476 rows; 519 MB size, 2,115,151 rows
NC: 262 MB size, 1,389,877 rows
OH: 497 MB size, 2,408,556 rows
NY: 512 MB size, 2,587,015 rows
VA: 111 MB size, 334,008 rows
WA: 205 MB size, 1,152,309 rows

Table Details: CorporateRegistrations_CA

<table>
<thead>
<tr>
<th>so_file_number</th>
<th>STRING</th>
</tr>
</thead>
<tbody>
<tr>
<td>corporation_number</td>
<td>INTEGER</td>
</tr>
<tr>
<td>corporation_status</td>
<td>STRING</td>
</tr>
<tr>
<td>corporation_classification</td>
<td>STRING</td>
</tr>
<tr>
<td>corporation_name</td>
<td>STRING</td>
</tr>
<tr>
<td>care_of_name</td>
<td>STRING</td>
</tr>
<tr>
<td>mail_address_line_1</td>
<td>STRING</td>
</tr>
<tr>
<td>mail_address_line_2</td>
<td>STRING</td>
</tr>
<tr>
<td>mail_address_city</td>
<td>STRING</td>
</tr>
<tr>
<td>mail_address_state_or_country</td>
<td>STRING</td>
</tr>
<tr>
<td>mail_address_zip_code</td>
<td>STRING</td>
</tr>
<tr>
<td>corporation_type</td>
<td>STRING</td>
</tr>
<tr>
<td>incorporation_date</td>
<td>DATE</td>
</tr>
<tr>
<td>so_file_date</td>
<td>DATE</td>
</tr>
<tr>
<td>term_expiration_date</td>
<td>DATE</td>
</tr>
<tr>
<td>chief_executive_officer_name</td>
<td>STRING</td>
</tr>
</tbody>
</table>
Secondary Dataset:
Occupational Employment Survey

Source: Bureau of Labor Statistics

Wages Table Sizes:
2015: 29.2 MB size, 473,717 rows
2016: 29.9 MB size, 484,390 rows
2017: 29.9 MB size, 484,390 rows
2018: 29.9 MB size, 485,211 rows

Geography Table Sizes:
2015: 340 KB size, 4,765 rows
2016: 357 KB size, 4,991 rows
2017: 357 KB size, 4,991 rows
2018: 357 KB size, 4,991 rows

Project Work:
- Imported files into BQ tables
- Milestone 9
Project Work:
- Merged corp. registration tables
- Merged wages tables
- Merged geography tables
- Normalized corp. name, city, state
Pay Gaps by Occupation:

Occupations which pay H1B workers higher than domestic workers

- COMPUTER SCIENCE TEACHER, POSTSECONDARY
- COMPUTER SUPPORT SPECIALISTS
- COMPUTER NETWORK SUPPORT SPECIALIST
- COMPUTER SOFTWARE ENGINEERS/ARCHITECTS

Occupations which pay H1B workers lower than domestic workers

- COMPUTER SYSTEMS ANALYSTS
- PROGRAMMER ANALYST
- COMPUTER PROGRAMMERS
- COMPUTER INFORMATION SYSTEMS MANAGERS