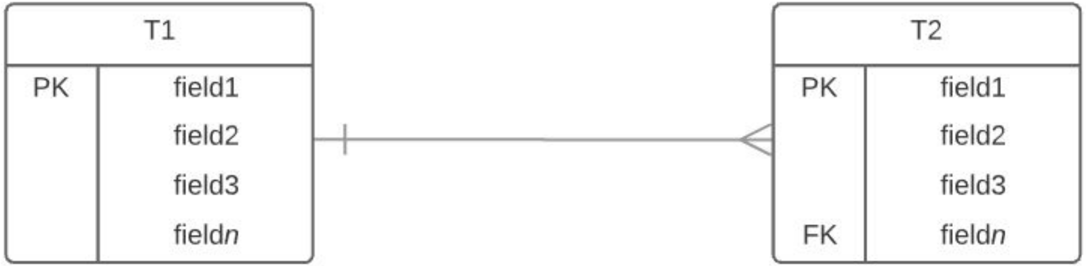
CS 327E Intro

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

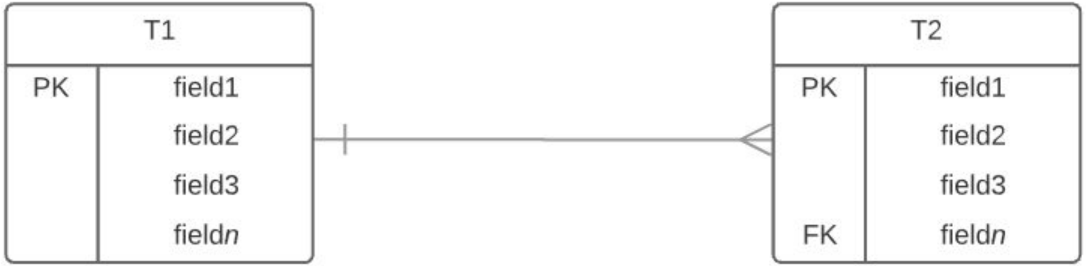
Table Relationship: One-to-Many (1:*m*)



T1	
PK	field1
	field2
	field3
	field*n*

T2	
PK	field1
	field2
	field3
FK	field*n*

Table Relationship: One-to-Many (1:*m*)



T1	
PK	field1
	field2
	field3
	field*n*

T2	
PK	field1
	field2
	field3
FK	field*n*

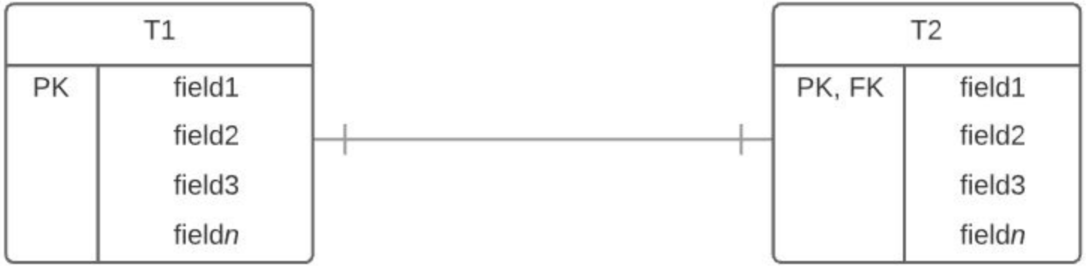
Author

<u>id</u>	name	section
1	Mary Tuma	news
2	Michael King	arts
3	Nina Hernandez	news
4	Sunil Kumar	music

Article

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

Table Relationship: One-to-One (1:1)



T1

PK

field1

field2

field3

field*n*

T2

PK, FK

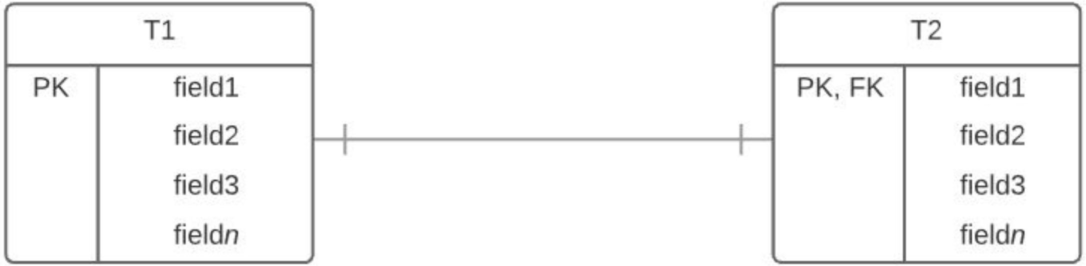
field1

field2

field3

field*n*

Table Relationship: One-to-One (1:1)



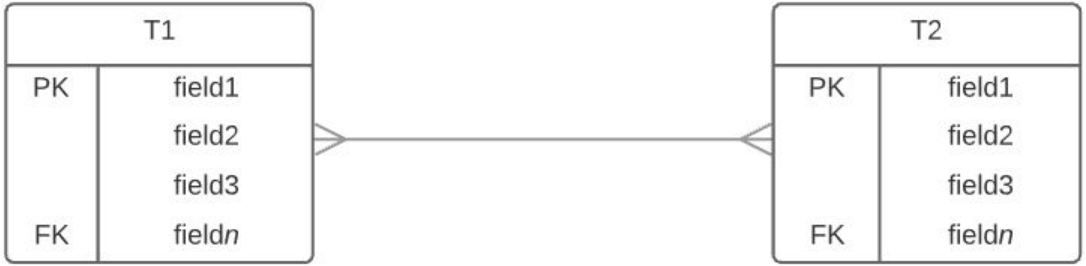
Article

id	title	date	*authid*
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

Article_Stats

id	clicks	likes	dislikes	comments
1	120	45	9	13
2	0	0	0	0
3	8	0	0	2
4	30	4	0	1
5	9	1	3	3

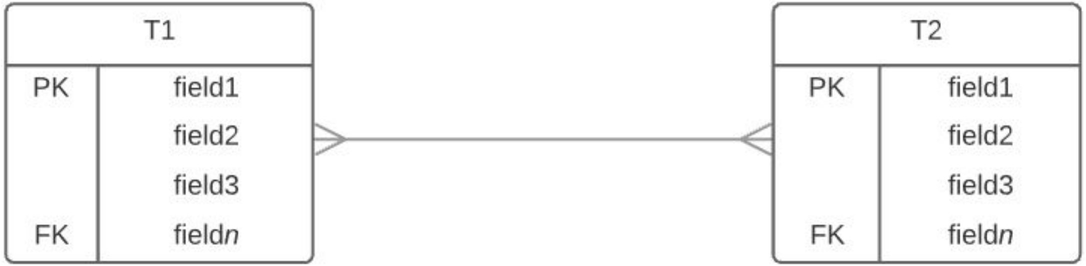
Table Relationship: Many-to-Many (*m*:*n*)



T1	
PK	field1
	field2
	field3
FK	field*n*

T2	
PK	field1
	field2
	field3
FK	field*n*

Table Relationship: Many-to-Many (m:n)



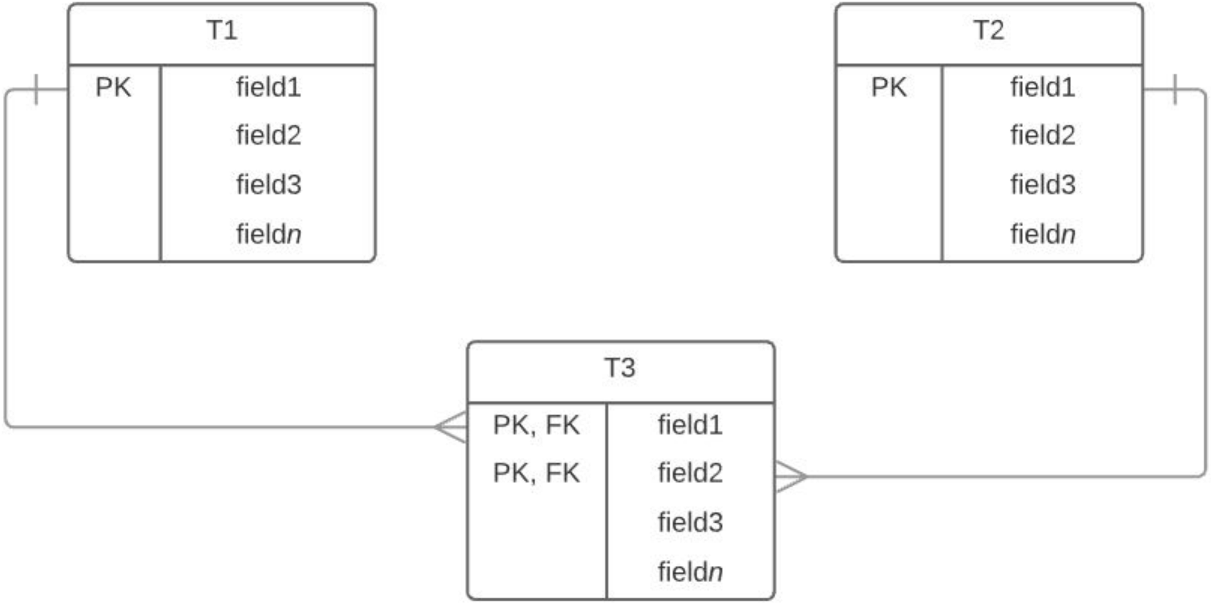
Tag

id	tag	*aids*
1	Politics	1, 2, 5
2	Austin	1, 2, 3, 4, 5
3	Mayor	3, 5
4	Business	1, 2, 5
6	Land Development	2
37	Animals	1

Article

id	title	date	*authid*	*tids*
1	Turmoil at the Zoo	2019-01-26	1	4, 37, 2
2	CodeNEXT's New Friend	2019-01-27	1	2, 6
3	Quote of the Week	2019-01-27	3	2, 3
4	SXSW News	2019-01-28	2	2, 40, 7
5	More from Steve Adler	2019-01-28	1	2, 3, 4

Representation of Many-to-Many (*m*:*n*)



T1

PK	field1
	field2
	field3
	field*n*

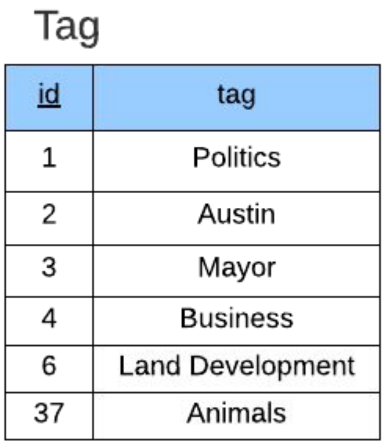
T2

PK	field1
	field2
	field3
	field*n*

T3

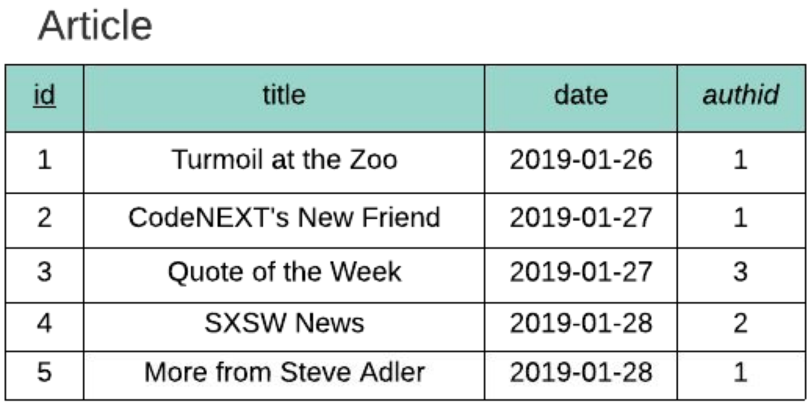
PK, FK	field1
PK, FK	field2
	field3
	field*n*

Table Relationship: Many-to-Many (*m*:*n*)



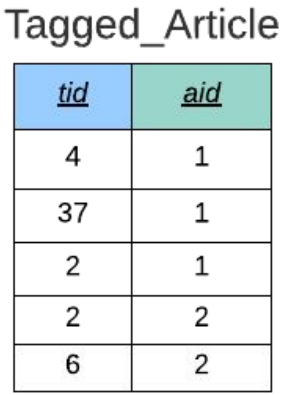
Tag

id	tag
1	Politics
2	Austin
3	Mayor
4	Business
6	Land Development
37	Animals



Article

id	title	date	*authid*
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1



Tagged_Article

tid	*aid*
4	1
37	1
2	1
2	2
6	2

iClicker Question

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

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

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

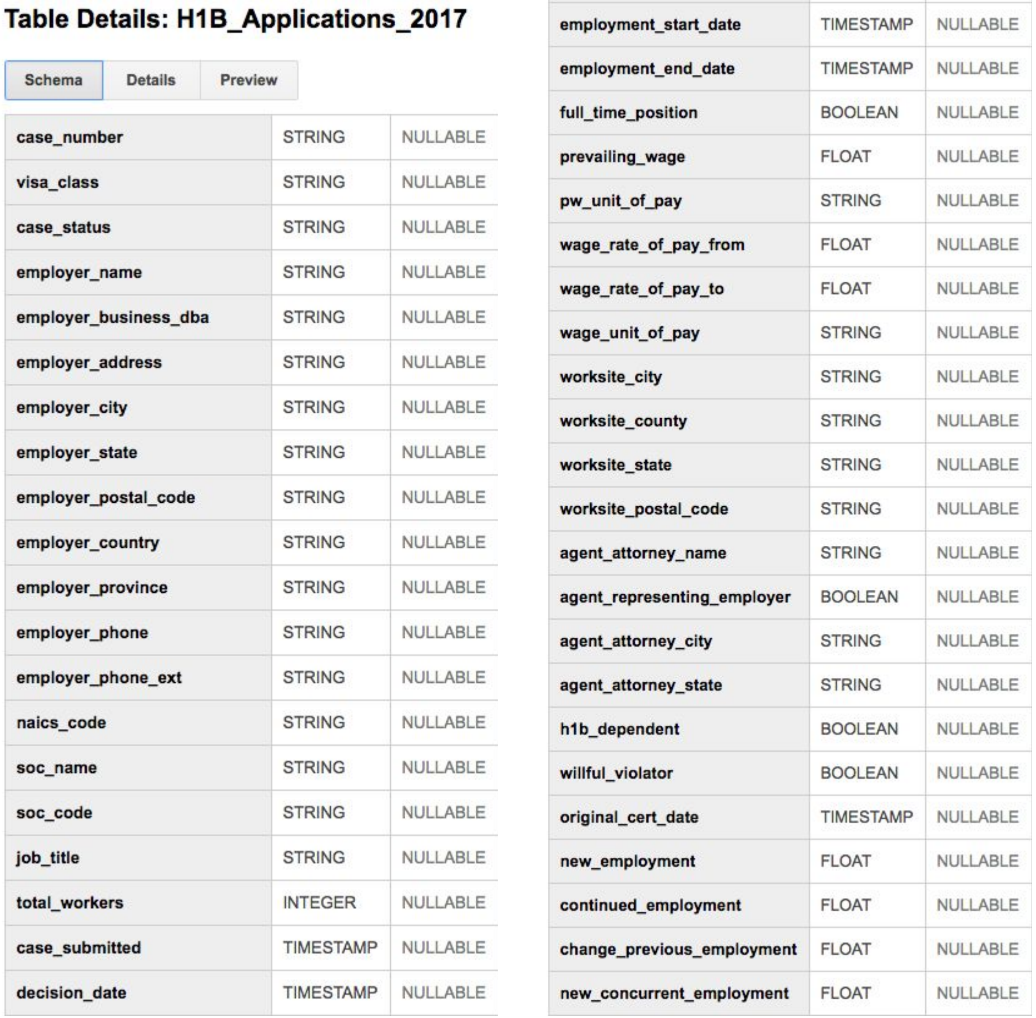
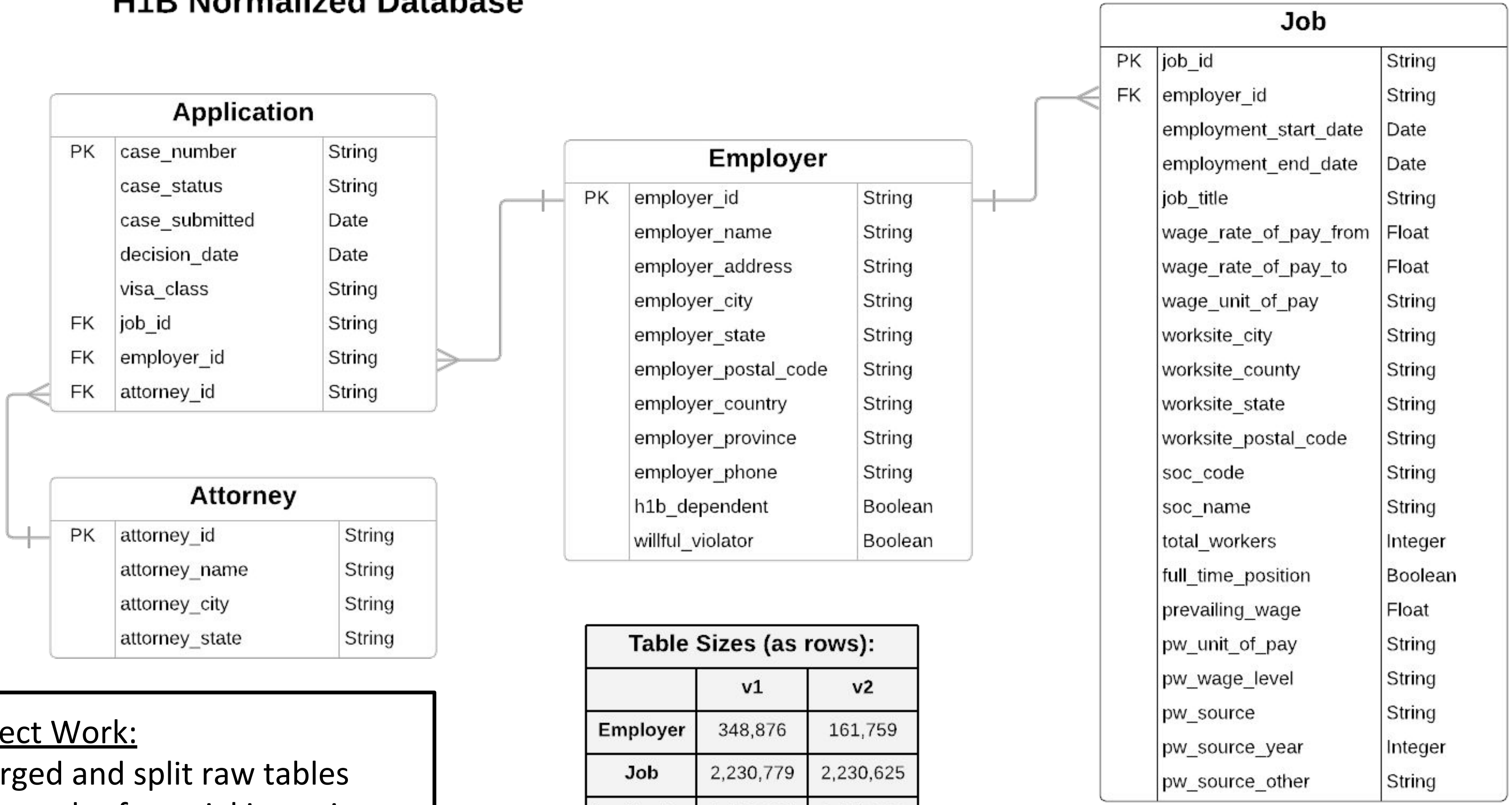


Table Details: H1B_Applications_2017

case_number	STRING	NULLABLE
visa_class	STRING	NULLABLE
case_status	STRING	NULLABLE
employer_name	STRING	NULLABLE
employer_business_dba	STRING	NULLABLE
employer_address	STRING	NULLABLE
employer_city	STRING	NULLABLE
employer_state	STRING	NULLABLE
employer_postal_code	STRING	NULLABLE
employer_country	STRING	NULLABLE
employer_province	STRING	NULLABLE
employer_phone	STRING	NULLABLE
employer_phone_ext	STRING	NULLABLE
naics_code	STRING	NULLABLE
soc_name	STRING	NULLABLE
soc_code	STRING	NULLABLE
job_title	STRING	NULLABLE
total_workers	INTEGER	NULLABLE
case_submitted	TIMESTAMP	NULLABLE
decision_date	TIMESTAMP	NULLABLE
employment_start_date	TIMESTAMP	NULLABLE
employment_end_date	TIMESTAMP	NULLABLE
full_time_position	BOOLEAN	NULLABLE
prevailing_wage	FLOAT	NULLABLE
pw_unit_of_pay	STRING	NULLABLE
wage_rate_of_pay_from	FLOAT	NULLABLE
wage_rate_of_pay_to	FLOAT	NULLABLE
wage_unit_of_pay	STRING	NULLABLE
worksite_city	STRING	NULLABLE
worksite_county	STRING	NULLABLE
worksite_state	STRING	NULLABLE
worksite_postal_code	STRING	NULLABLE
agent_attorney_name	STRING	NULLABLE
agent_representing_employer	BOOLEAN	NULLABLE
agent_attorney_city	STRING	NULLABLE
agent_attorney_state	STRING	NULLABLE
h1b_dependent	BOOLEAN	NULLABLE
willful_violator	BOOLEAN	NULLABLE
original_cert_date	TIMESTAMP	NULLABLE
new_employment	FLOAT	NULLABLE
continued_employment	FLOAT	NULLABLE
change_previous_employment	FLOAT	NULLABLE
new_concurrent_employment	FLOAT	NULLABLE

H1B Normalized Database



Application

PK	case_number	String
	case_status	String
	case_submitted	Date
	decision_date	Date
	visa_class	String
FK	job_id	String
FK	employer_id	String
FK	attorney_id	String

Attorney

PK	attorney_id	String
	attorney_name	String
	attorney_city	String
	attorney_state	String

Employer

PK	employer_id	String
	employer_name	String
	employer_address	String
	employer_city	String
	employer_state	String
	employer_postal_code	String
	employer_country	String
	employer_province	String
	employer_phone	String
	h1b_dependent	Boolean
	willful_violator	Boolean

Job

PK	job_id	String
FK	employer_id	String
	employment_start_date	Date
	employment_end_date	Date
	job_title	String
	wage_rate_of_pay_from	Float
	wage_rate_of_pay_to	Float
	wage_unit_of_pay	String
	worksite_city	String
	worksite_county	String
	worksite_state	String
	worksite_postal_code	String
	soc_code	String
	soc_name	String
	total_workers	Integer
	full_time_position	Boolean
	prevailing_wage	Float
	pw_unit_of_pay	String
	pw_wage_level	String
	pw_source	String
	pw_source_year	Integer
	pw_source_other	String

Table Sizes (as rows):

	v1	v2
Employer	348,876	161,759
Job	2,230,779	2,230,625
Application	2,633,426	2,633,156
Attorney	19,861	N/A

Project Work:

-Merged and split raw tables

-Enforced referential integrity

-Removed duplicate records

-Milestones 4, 5, 6

Secondary Dataset: Corporate Registrations

<u>Source:</u>
Secretary of States

<u>Table Sizes:</u>
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

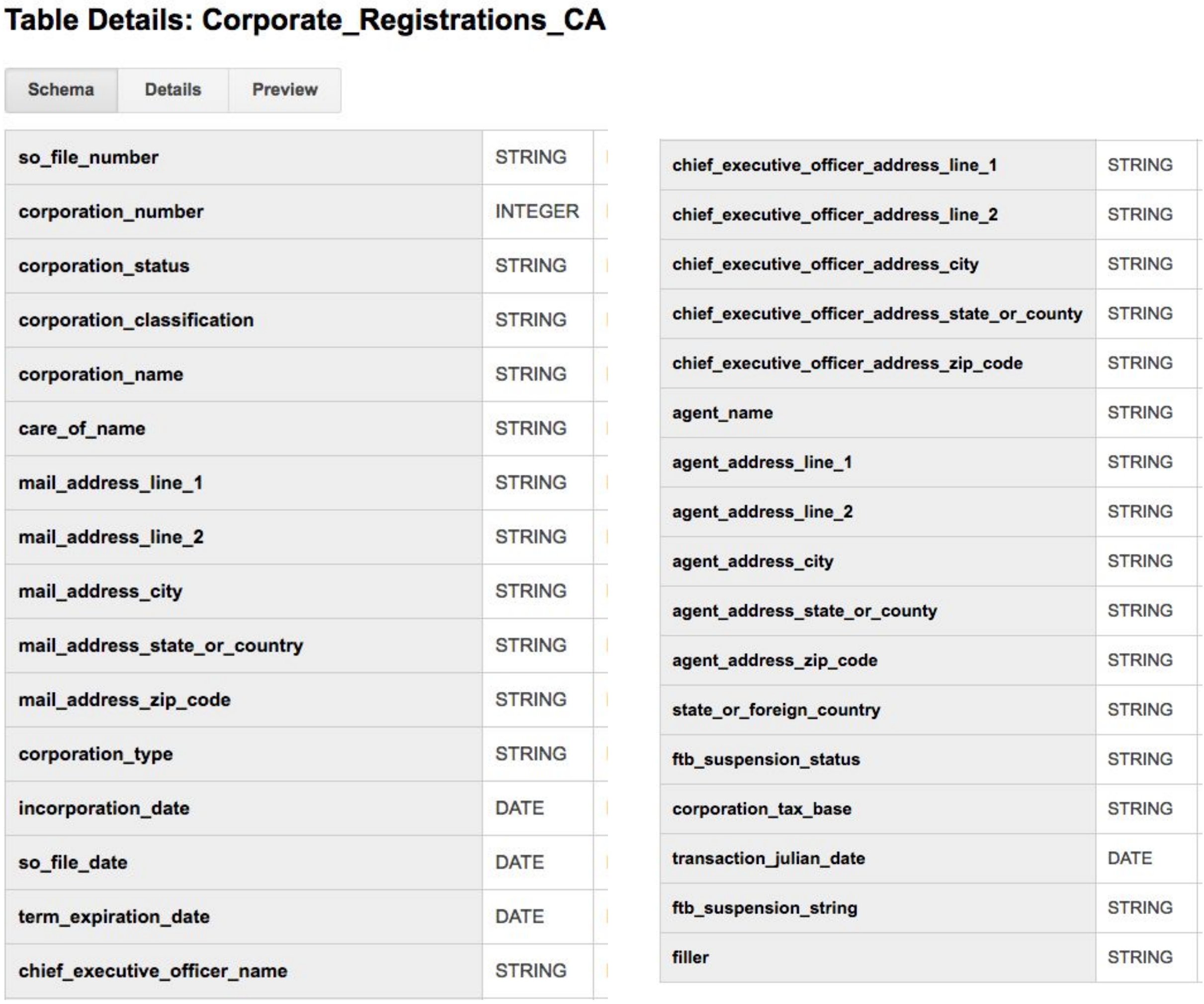


Table Details: Corporate_Registrations_CA

so_file_number	STRING
corporation_number	INTEGER
corporation_status	STRING
corporation_classification	STRING
corporation_name	STRING
care_of_name	STRING
mail_address_line_1	STRING
mail_address_line_2	STRING
mail_address_city	STRING
mail_address_state_or_country	STRING
mail_address_zip_code	STRING
corporation_type	STRING
incorporation_date	DATE
so_file_date	DATE
term_expiration_date	DATE
chief_executive_officer_name	STRING
chief_executive_officer_address_line_1	STRING
chief_executive_officer_address_line_2	STRING
chief_executive_officer_address_city	STRING
chief_executive_officer_address_state_or_county	STRING
chief_executive_officer_address_zip_code	STRING
agent_name	STRING
agent_address_line_1	STRING
agent_address_line_2	STRING
agent_address_city	STRING
agent_address_state_or_county	STRING
agent_address_zip_code	STRING
state_or_foreign_country	STRING
ftb_suspension_status	STRING
corporation_tax_base	STRING
transaction_julian_date	DATE
ftb_suspension_string	STRING
filler	STRING

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

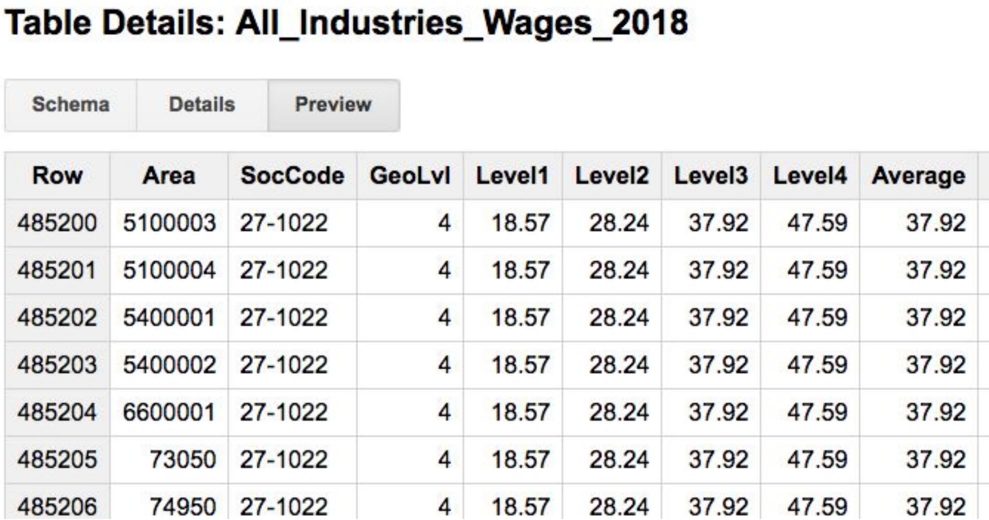


Table Details: All_Industries_Wages_2018

Row	Area	SocCode	GeoLvl	Level1	Level2	Level3	Level4	Average
485200	5100003	27-1022	4	18.57	28.24	37.92	47.59	37.92
485201	5100004	27-1022	4	18.57	28.24	37.92	47.59	37.92
485202	5400001	27-1022	4	18.57	28.24	37.92	47.59	37.92
485203	5400002	27-1022	4	18.57	28.24	37.92	47.59	37.92
485204	6600001	27-1022	4	18.57	28.24	37.92	47.59	37.92
485205	73050	27-1022	4	18.57	28.24	37.92	47.59	37.92
485206	74950	27-1022	4	18.57	28.24	37.92	47.59	37.92

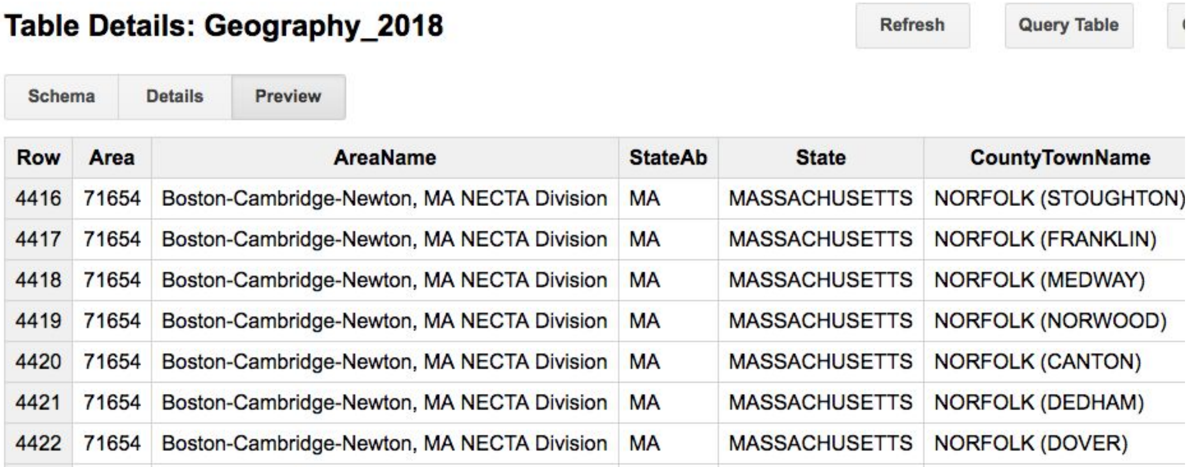
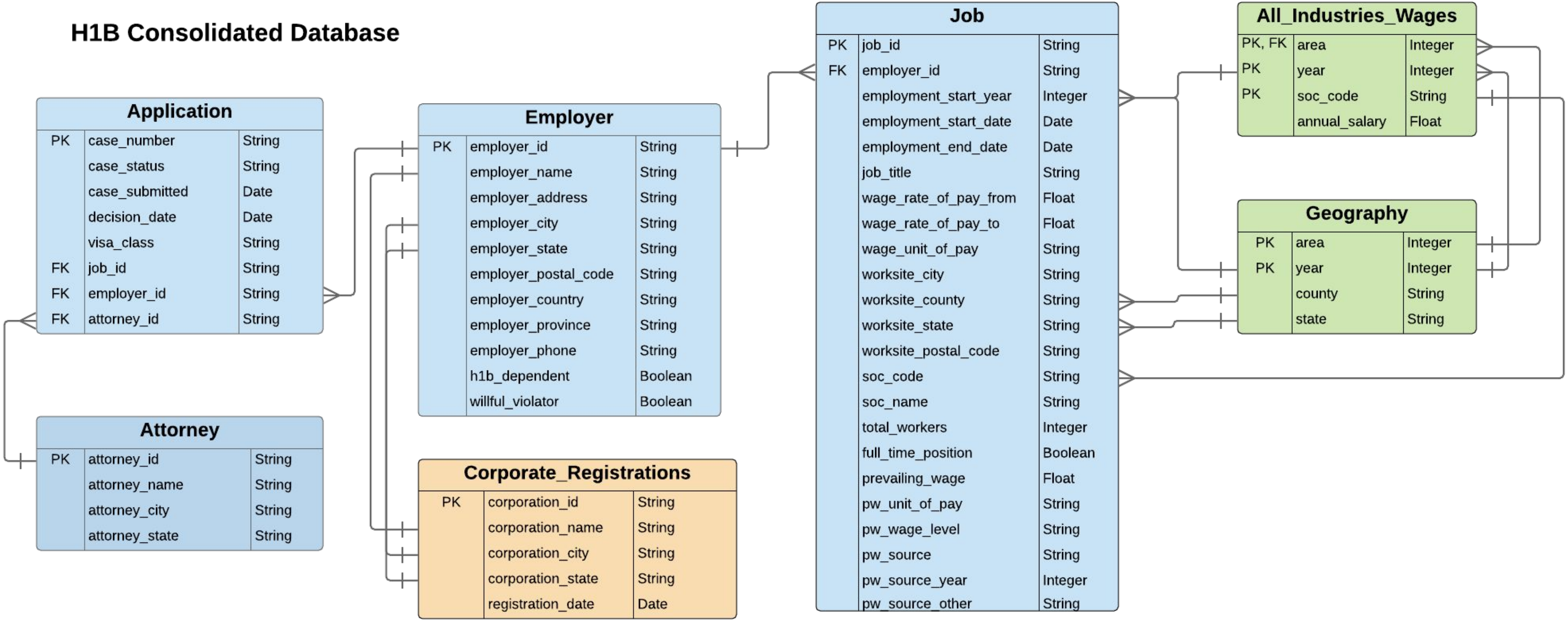


Table Details: Geography_2018

Row	Area	AreaName	StateAb	State	CountyTownName
4416	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (STOUGHTON)
4417	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (FRANKLIN)
4418	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (MEDWAY)
4419	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (NORWOOD)
4420	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (CANTON)
4421	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (DEDHAM)
4422	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (DOVER)

H1B Consolidated Database



Application

PK	case_number	String
	case_status	String
	case_submitted	Date
	decision_date	Date
	visa_class	String
FK	job_id	String
FK	employer_id	String
FK	attorney_id	String

Attorney

PK	attorney_id	String
	attorney_name	String
	attorney_city	String
	attorney_state	String

Employer

PK	employer_id	String
	employer_name	String
	employer_address	String
	employer_city	String
	employer_state	String
	employer_postal_code	String
	employer_country	String
	employer_province	String
	employer_phone	String
	h1b_dependent	Boolean
	willful_violator	Boolean

Corporate_Registrations

PK	corporation_id	String
	corporation_name	String
	corporation_city	String
	corporation_state	String
	registration_date	Date

Job

PK	job_id	String
FK	employer_id	String
	employment_start_year	Integer
	employment_start_date	Date
	employment_end_date	Date
	job_title	String
	wage_rate_of_pay_from	Float
	wage_rate_of_pay_to	Float
	wage_unit_of_pay	String
	worksite_city	String
	worksite_county	String
	worksite_state	String
	worksite_postal_code	String
	soc_code	String
	soc_name	String
	total_workers	Integer
	full_time_position	Boolean
	prevailing_wage	Float
	pw_unit_of_pay	String
	pw_wage_level	String
	pw_source	String
	pw_source_year	Integer
	pw_source_other	String

All_Industries_Wages

PK, FK	area	Integer
PK	year	Integer
PK	soc_code	String
	annual_salary	Float

Geography

PK	area	Integer
PK	year	Integer
	county	String
	state	String

Project Work:

-Merged corp. registration tables

-Merged wages tables

-Merged geography tables

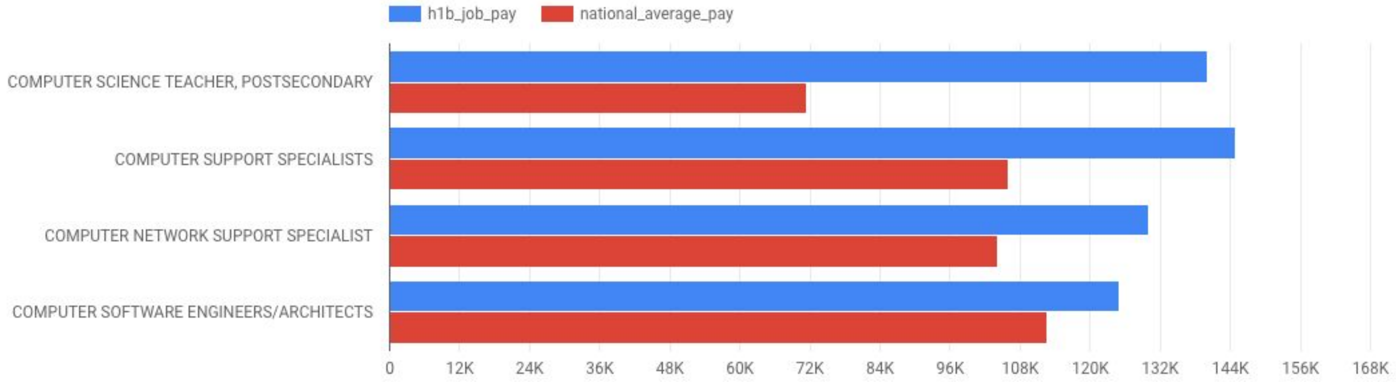
-Normalized corp. name, city, state

Milestones 10, 11, 12

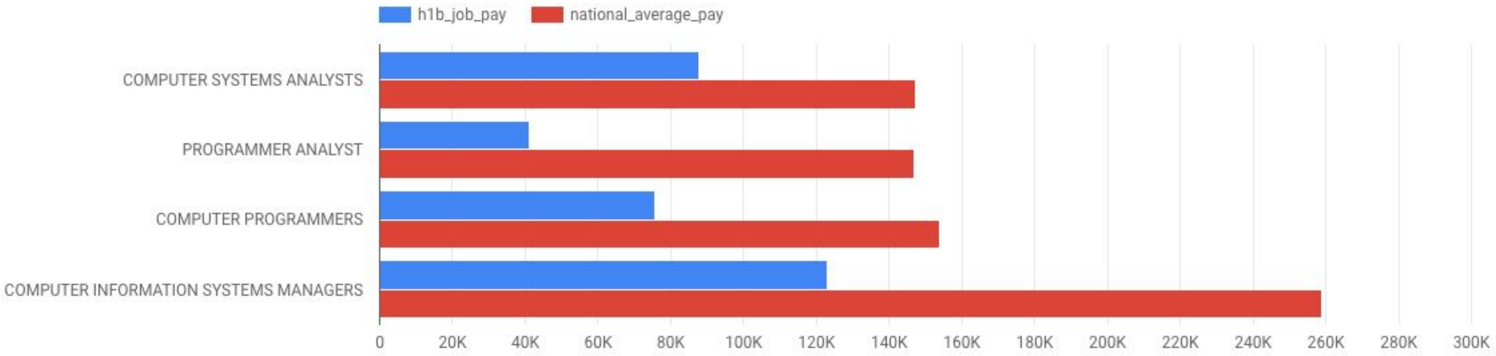
Sample Reports

Pay Gaps by Occupation:

Occupations which pay H1B workers *higher* than domestic workers



Occupations which pay H1B workers *lower* than domestic workers


🟦 h1b_job_pay
🟥 national_average_pay
COMPUTER SYSTEMS ANALYSTS
PROGRAMMER ANALYST
COMPUTER PROGRAMMERS
COMPUTER INFORMATION SYSTEMS MANAGERS
20K
40K
60K
80K
100K
120K
140K
160K
180K
200K
220K
240K
260K
280K
300K