CS 327E Class 11

December 3, 2018

Announcements

- Group presentations start next Monday!
- Presentation schedule is posted
- CIS is open, take survey by next Monday. **Your voice matters :)**

1) Which representation of a `Sales_Order` uses a **denormalized** structure?

A.
```
Sales_Order(order_id:STRING, cust_id:INTEGER,
cust_name:STRING, cust_email:STRING, timestamp:DATETIME,
location:STRING, purchase_items:RECORD[sku:INTEGER,
description:STRING, quantity:INTEGER, price:FLOAT]);
```

B.
```
Sales_Order(order_id:STRING, cust_id:INTEGER,
            timestamp:DATETIME, location:STRING);
Purchase_Item(sku:INTEGER, description:STRING,
            quantity:INTEGER, price:FLOAT, order_id:STRING);
Customer(cust_id:INTEGER,cust_name:STRING,
          cust_email:STRING);
```

2) What is the main advantage of **denormalized** structures?

A. They reduce query time, especially on large tables.
B. They require less storage than normalized structures.
C. They are more performant for update and delete operations.

3) What is the main disadvantage of **denormalized** structures?

A. They can only store JSON data in a table.
B. They make it more challenging to maintain data integrity.
C. They can’t express *m:n* relationships between tables.

4) Which BigQuery operator lets you access individual elements of a repeated field?

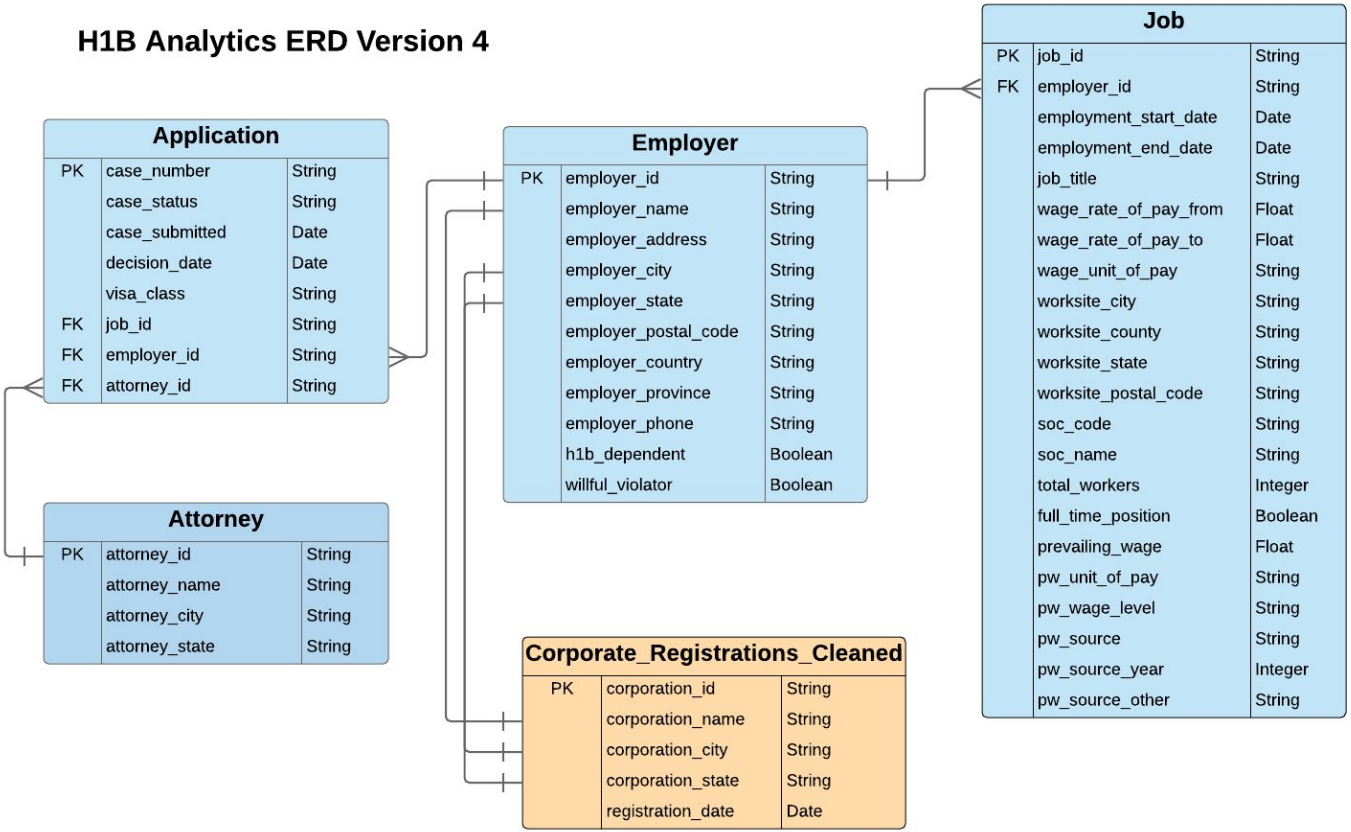
A. ARRAY
B. FLATTEN
C. UNNEST

5) What does the **view switching** technique require?

A. Masking duplicate records of a table.
B. Maintaining two copies of data during update cycles.
C. Capturing each insert, update, and delete operation in a separate history table.

Case Study: Part 3

H1B Analytics ERD Version 4



Notes:

New Source Tables:
sec_of_state.Corporate_Registrations_Merged.

New Target Table:
-sec_of_state.Corporate_Registrations_Cleaned.
-generated from Beam pipeline.

Changes since previous version:
- removed punctuation marks and suffixes from corporation_name.
- performed simple validation of corporation_city.
- cross-dataset join returns **12,856** results (instead of only 804 results).

Number of Rows		
	v1	**v2**
Corporate_Registrations	16,379,107	16,321,932
Employer	348,876	161,759
v_Tech_Employer_13_States	29,658	

Third Dataset

Table Details: All_Industries_Wages_2018

Schema | Details | Preview

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

Wages Table Details:
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 Details:
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

Table Details: Geography_2018

Refresh | Query Table | Co

Schema | Details | Preview

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)

SQL Transforms

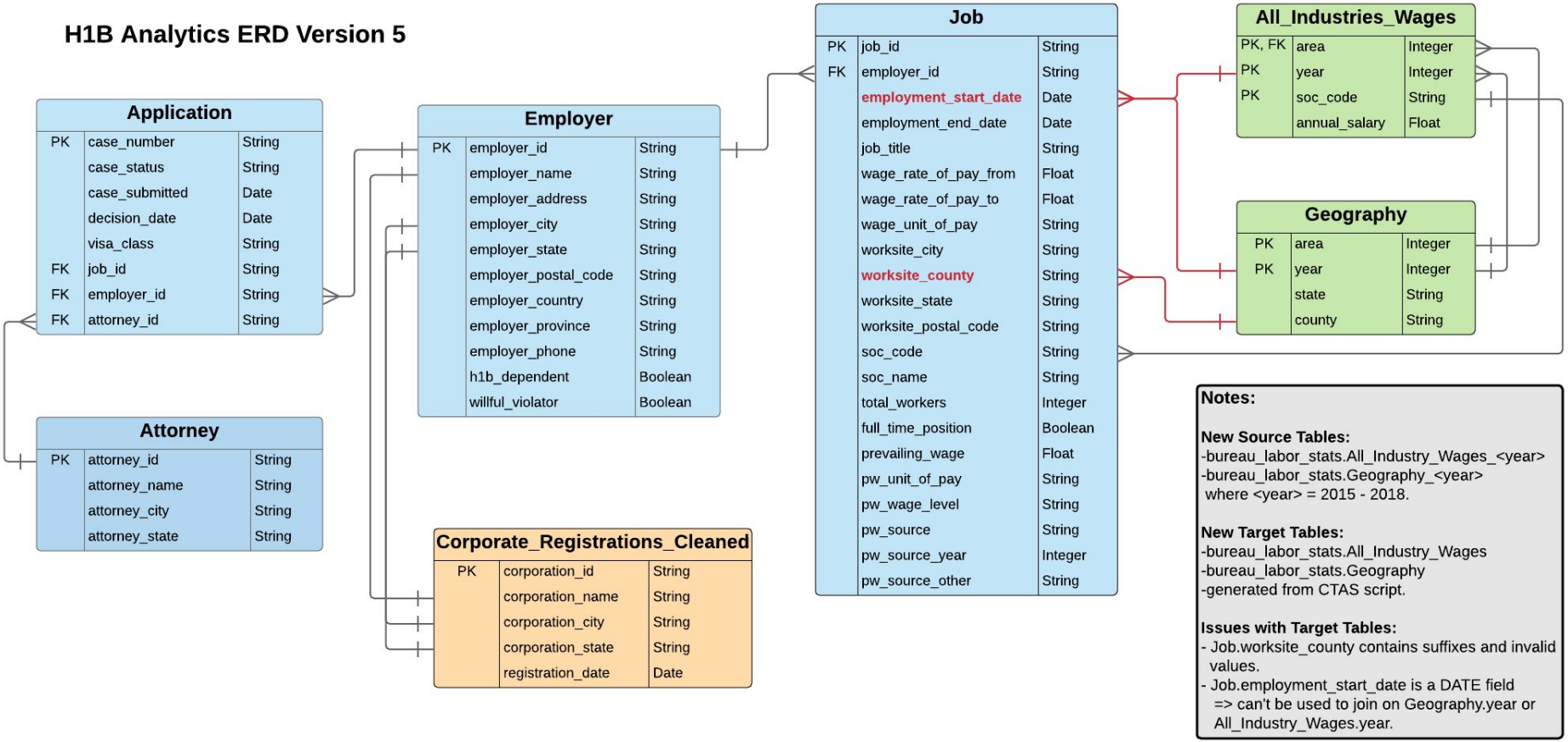
- Unions the yearly staging tables for Wages and Geography (2015 - 2018)
- Excludes all unwanted fields from result tables
- Calculates annual salaries from average wages per occupation and geo area

```
create table bureau_labor_stats.All_Industries_Wages
(
    area INT64,
    year INT64,
    soc_code STRING,
    annual_salary FLOAT64,
    empty_date DATE
)
PARTITION BY empty_date
CLUSTER BY year;

insert into bureau_labor_stats.All_Industries_Wages (area, year, soc_code, annual_salary)
select Area, 2015, SocCode,
(case when Average < 300 then round(((Average*8)*365), 2)
 when Average > 15000 then round(Average, 2)
 else NULL end)
from bureau_labor_stats.All_Industries_Wages_2015;

insert into bureau_labor_stats.All_Industries_Wages (area, year, soc_code, annual_salary)
select Area, 2016, SocCode,
(case when Average < 300 then round(((Average*8)*365), 2)
 when Average > 15000 then round(Average, 2)
 else NULL end)
from bureau_labor_stats.All_Industries_Wages_2016;
```

```
create table bureau_labor_stats.Geography
(
    area INT64,
    year INT64,
    state STRING,
    county STRING,
    empty_date DATE
)
PARTITION BY empty_date
CLUSTER BY year;

insert into bureau_labor_stats.Geography (area, year, state, county)
select Area, 2015, StateAb, CountyTownName
from bureau_labor_stats.Geography_2015;

insert into bureau_labor_stats.Geography (area, year, state, county)
select Area, 2016, StateAb, CountyTownName
from bureau_labor_stats.Geography_2016;

insert into bureau_labor_stats.Geography (area, year, state, county)
select Area, 2017, StateAb, CountyTownName
from bureau_labor_stats.Geography_2017;
```

Source File: https://github.com/shirleycohen/h1b_analytics/blob/master/bureau_labor_stats_ctas.sql

H1B Analytics ERD Version 5



Notes:

New Source Tables:
-bureau_labor_stats.All_Industry_Wages_<year>
-bureau_labor_stats.Geography_<year>
where <year> = 2015 - 2018.

New Target Tables:
-bureau_labor_stats.All_Industry_Wages
-bureau_labor_stats.Geography
-generated from CTAS script.

Issues with Target Tables:
- Job.worksite_county contains suffixes and invalid values.
- Job.employment_start_date is a DATE field => can't be used to join on Geography.year or All_Industry_Wages.year.

Data Integrity Issues

New Query

Query Editor | UDF Editor

```
1  select * from h1b_split.Job where worksite_county like '% COUNTY'
```

SQL

Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete.

RUN QUERY | Save Query | Save View | Format Query | Schedule Query | Show Options

Query complete (4.6s elapsed, 563 MB processed)

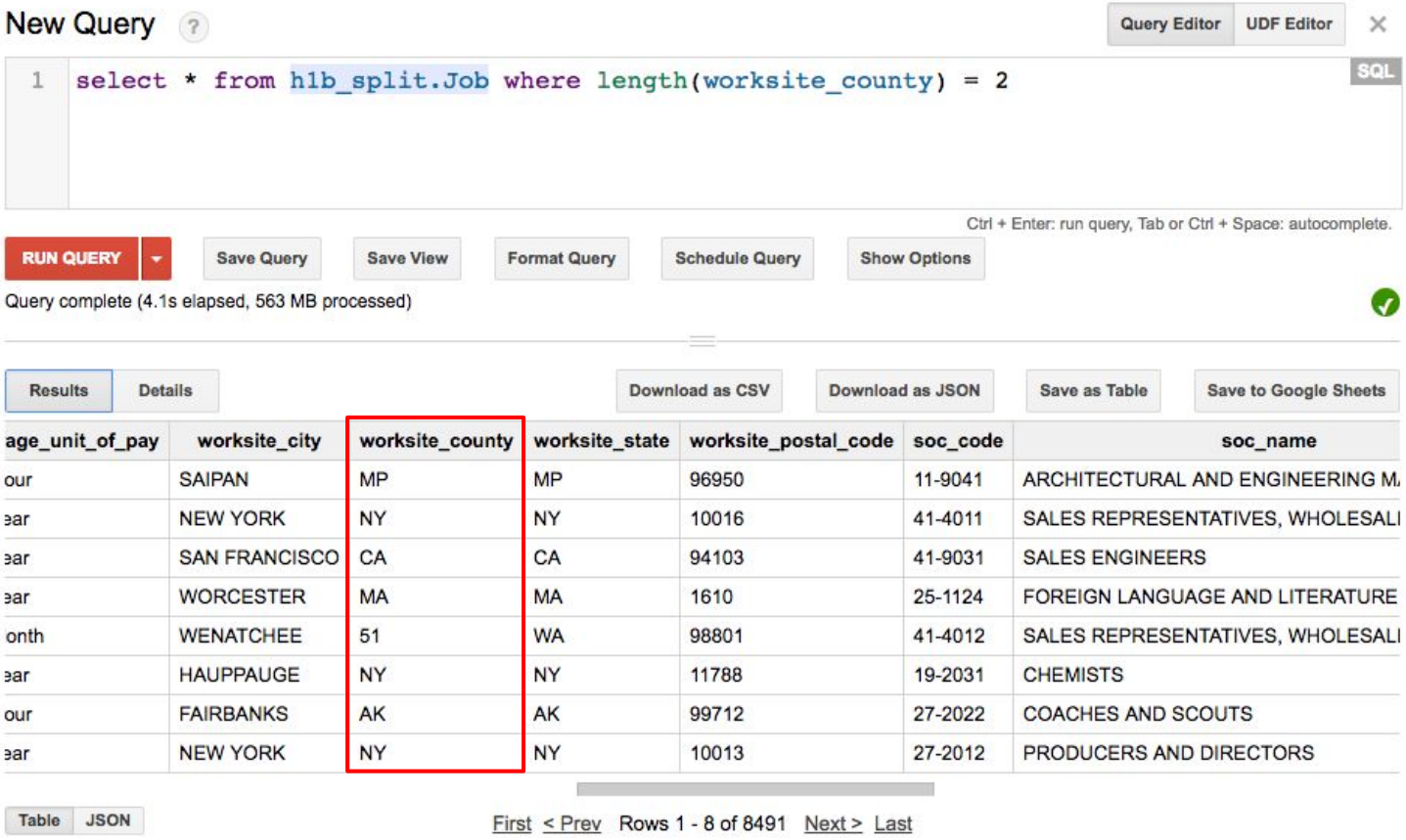
Results | Details | Download as CSV | Download as JSON | Save as Table | Save to Google Sheets

worksite_city	worksite_county	worksite_state	worksite_postal_code	soc_code	soc_name
HOUSTON	HARRIS COUNTY	TX	77027	19-4061	SOCIAL SCIENCE RESEARCH ASSISTANT
BROOKLYN	KINGS COUNTY	NY	11201	27-3031	PUBLIC RELATIONS SPECIALISTS
LAGUNA BEACH	ORANGE COUNTY	CA	92651	17-1012	LANDSCAPE ARCHITECTS
HUNTINGTON BEACH	ORANGE COUNTY	CA	92648	23-1012	JUDICIAL LAW CLERKS
SEATTLE	KING COUNTY	WA	98125	19-3092	GEOGRAPHERS
ONTARIO	SAN BERNARDINO COUNTY	CA	91761	13-2031	BUDGET ANALYSTS
MIAMI LAKES	DADE COUNTY	FL	33014	23-2099	LEGAL SUPPORT WORKERS, ALL OTHER
THE WOODLANDS	MONTGOMERY COUNTY	TX	77380	13-1051	COST ESTIMATORS

Table | JSON

First < Prev Rows 1 - 8 of 48451 Next > Last

Data Integrity Issues



New Query

Query Editor | UDF Editor

```
1  select * from h1b_split.Job where length(worksite_county) = 2
```

SQL

Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete.

RUN QUERY | Save Query | Save View | Format Query | Schedule Query | Show Options

Query complete (4.1s elapsed, 563 MB processed)

Results | Details | Download as CSV | Download as JSON | Save as Table | Save to Google Sheets

age_unit_of_pay	worksite_city	worksite_county	worksite_state	worksite_postal_code	soc_code	soc_name
our	SAIPAN	MP	MP	96950	11-9041	ARCHITECTURAL AND ENGINEERING M
ear	NEW YORK	NY	NY	10016	41-4011	SALES REPRESENTATIVES, WHOLESALI
ear	SAN FRANCISCO	CA	CA	94103	41-9031	SALES ENGINEERS
ear	WORCESTER	MA	MA	1610	25-1124	FOREIGN LANGUAGE AND LITERATURE
onth	WENATCHEE	51	WA	98801	41-4012	SALES REPRESENTATIVES, WHOLESALI
ear	HAUPPAUGE	NY	NY	11788	19-2031	CHEMISTS
our	FAIRBANKS	AK	AK	99712	27-2022	COACHES AND SCOUTS
ear	NEW YORK	NY	NY	10013	27-2012	PRODUCERS AND DIRECTORS

Table | JSON

First < Prev Rows 1 - 8 of 8491 Next > Last

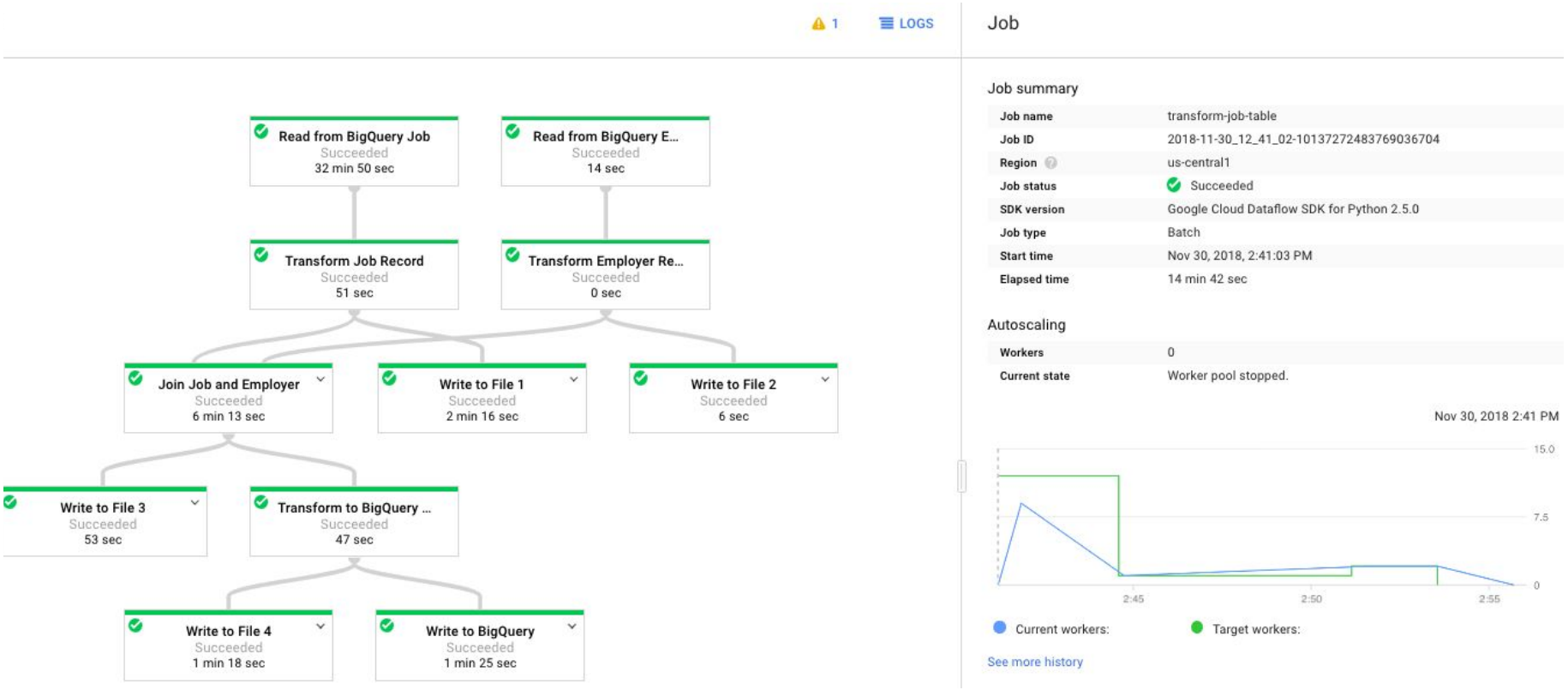
Beam Transforms

```
216 with beam.Pipeline('DataflowRunner', options=opts) as p:
217
218     job_query_str = 'SELECT *, EXTRACT(YEAR FROM employment_start_date) AS employment_start_year ' \
219                     'FROM `h1b_split.Job_Temp` ORDER BY employer_name'
220
221     emp_query_str = 'SELECT employer_id, employer_name, employer_city FROM ' \
222                     '`h1b_split.Employer` ORDER BY employer_name'
223
224     job_query_results = p | 'Read from BigQuery Job' >> beam.io.Read(beam.io.BigQuerySource(query=job_query_str, use_standard_sql=True))
225     emp_query_results = p | 'Read from BigQuery Employer' >> beam.io.Read(beam.io.BigQuerySource(query=emp_query_str, use_standard_sql=True))
226
227     # apply ParDo to the Job records
228     job_tuple_pcoll = job_query_results | 'Transform Job Record' >> beam.ParDo(TransformJobRecord())
229     emp_tuple_pcoll = emp_query_results | 'Transform Employer Record' >> beam.ParDo(TransformEmployerRecord())
```

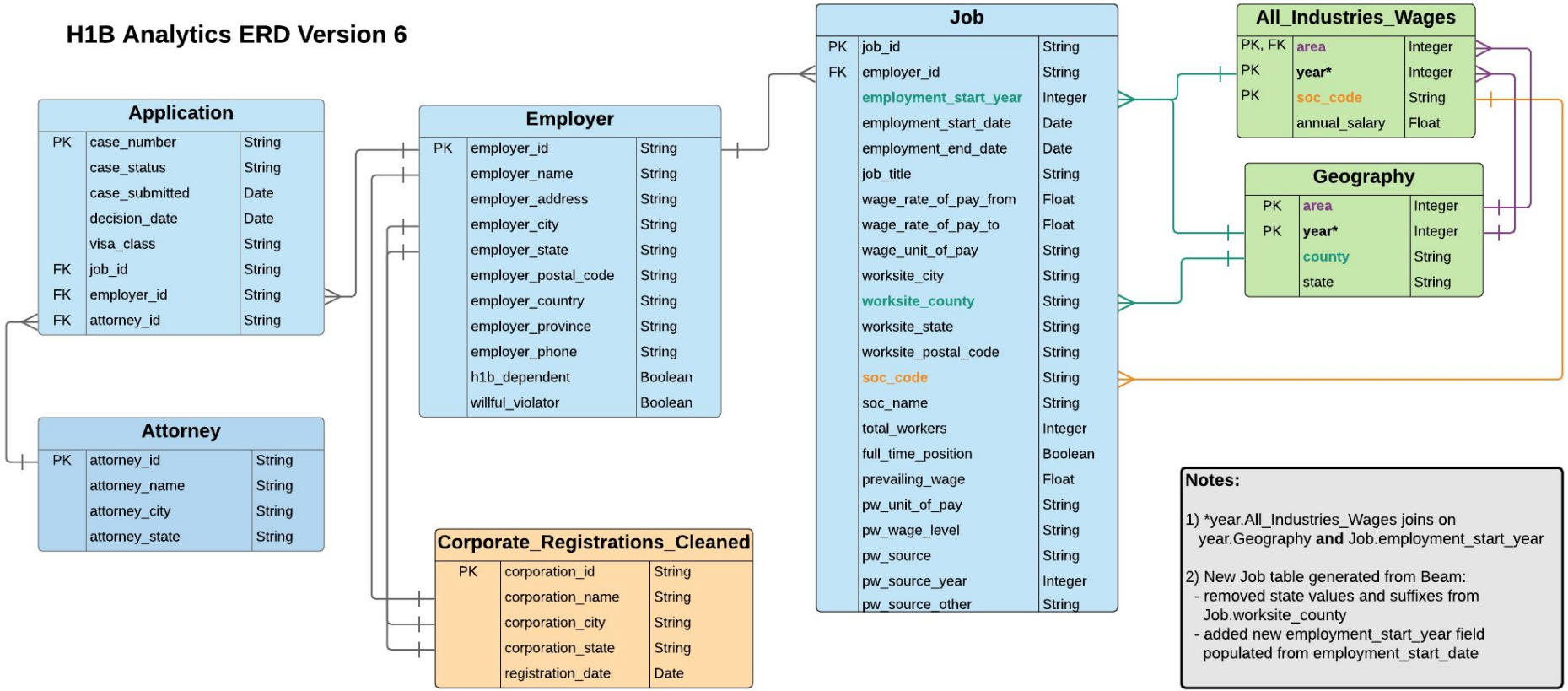
```
# remove suffix from worksite_county
if worksite_county != None:
    worksite_county = worksite_county.replace(' COUNTY', '')
    job_record['worksite_county'] = worksite_county

# county and state should not be equal to each other
if worksite_county == worksite_state:
    worksite_county = None
    job_record.pop('worksite_county')
if worksite_county != None and worksite_county.isdigit():
    worksite_county = None
    job_record.pop('worksite_county')
```

Source File: https://github.com/shirleycohen/h1b_analytics/blob/master/transform_job_table_cluster.py

Dataflow Job



H1B Analytics ERD Version 6



Cross-Dataset Query

New Query

Query Editor | UDF Editor

SQL

```
SELECT j.job_id, j.worksite_county, g.county, w.area, j.worksite_state, g.state,
 j.soc_code, j.soc_name, j.employment_start_year as year, j.wage_rate_of_pay_from as job_salary,
 w.annual_salary as national_salary
FROM h1b_split.v_Tech_Job j JOIN bureau_labor_stats.Geography g
 ON (j.worksite_county = g.county AND j.employment_start_year = g.year)
 JOIN bureau_labor_stats.All_Industries_Wages w
 ON (g.area = w.area and g.year = w.year)
WHERE j.soc_code = w.soc_code AND j.employment_start_year = w.year
ORDER BY j.job_id
```

Standard SQL Dialect

Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete.

RUN QUERY | Save Query | Save View | Format Query | Schedule Query | Show Options

Results | Details

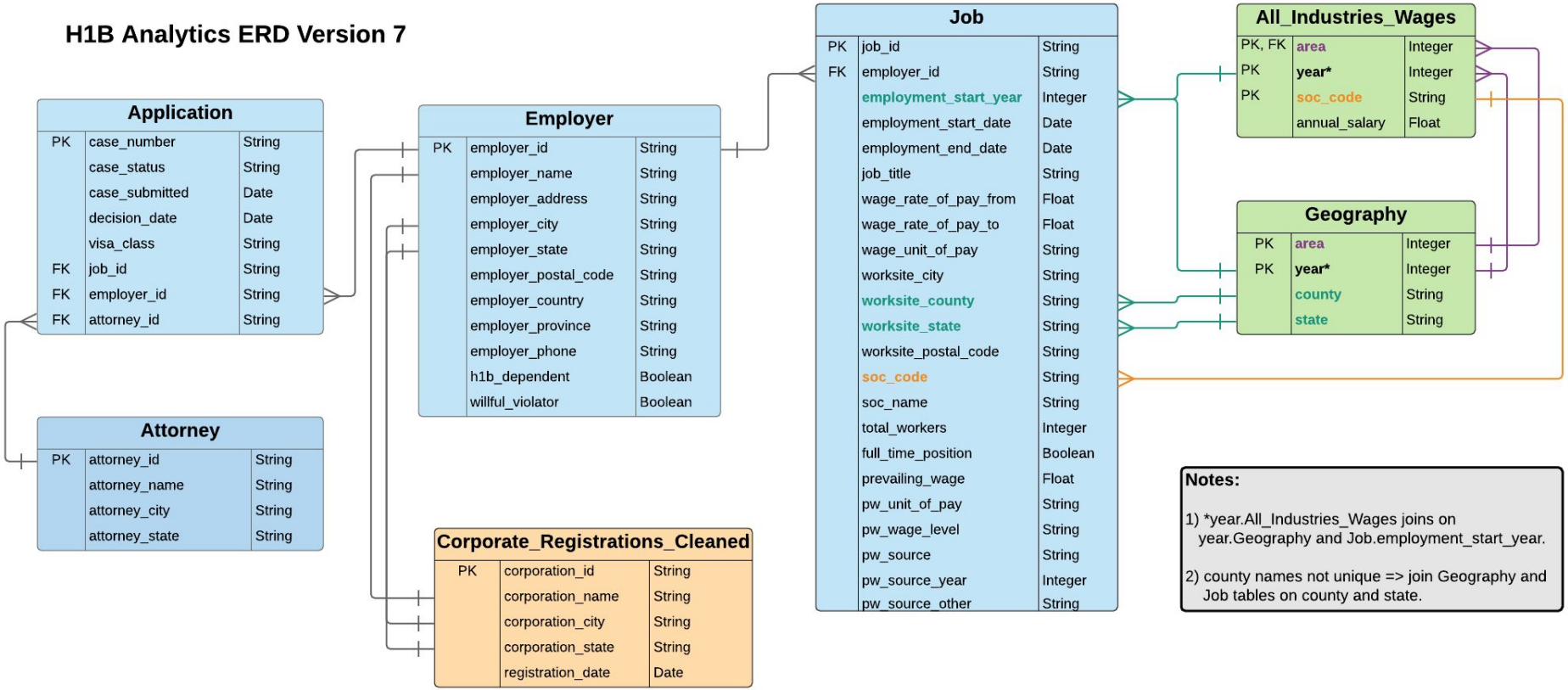
Download as CSV | Download as JSON | Save as Table | Save to Google Sheets

Row	job_id	worksite_county	county	area	worksite_state	state	soc_code	soc_name	year	job_salary	national_salary
1	0000035e-b44f-4785-b2e4-a8d10e4e8857	TOMPKINS	TOMPKINS	27060	NY	NY	15-1131	COMPUTER PROGRAMMERS	2016	64000.0	109850.4
2	000013d8-82cd-4ba5-8d17-cbab48d8d5b6	KING	KING	42644	WA	WA	15-1121	COMPUTER SYSTEMS ANALYSTS	2017	84000.0	139138.0
3	000013d8-82cd-4ba5-8d17-cbab48d8d5b6	KING	KING	4800001	WA	TX	15-1121	COMPUTER SYSTEMS ANALYSTS	2017	84000.0	95805.2
4	00001e51-1306-46ff-8fed-e8173fb62d54	MARICOPA	MARICOPA	38060	AZ	AZ	15-1132	SOFTWARE DEVELOPERS, APPLICATIONS	2017	92435.0	132655.6
5	0000287a-c470-4130-858e-d428865b70ce	BENTON	BENTON	22220	AR	AR	15-1131	COMPUTER PROGRAMMERS	2018	65187.0	92534.8
6	0000287a-c470-4130-858e-d428865b70ce	BENTON	BENTON	29200	AR	IN	15-1131	COMPUTER PROGRAMMERS	2018	65187.0	91658.8
7	0000287a-c470-4130-858e-d428865b70ce	BENTON	BENTON	41060	AR	MN	15-1131	COMPUTER PROGRAMMERS	2018	65187.0	127487.2
8	0000287a-c470-4130-858e-d428865b70ce	BENTON	BENTON	2900001	AR	MO	15-1131	COMPUTER PROGRAMMERS	2018	65187.0	100185.2

Table | JSON

First < Prev Rows 1 - 8 of 4502905 Next > Last

H1B Analytics ERD Version 7



Revised Cross-Dataset Query

New Query

Query Editor | UDF Editor

```
SELECT j.job_id, j.worksite_county, g.county, w.area, j.worksite_state, g.state,
 j.soc_code, j.soc_name, j.employment_start_year as year, j.wage_rate_of_pay_from as job_salary,
 w.annual_salary as national_salary
FROM h1b_split.v_Tech_Job_j JOIN bureau_labor_stats.Geography g
 ON (j.worksite_county = g.county AND j.worksite_state = g.state AND j.employment_start_year = g.year)
 JOIN bureau_labor_stats.All_Industries_Wages w
 ON (g.area = w.area and g.year = w.year)
WHERE j.soc_code = w.soc_code AND j.employment_start_year = w.year
ORDER BY j.job_id
```

Standard SQL Dialect

Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete

RUN QUERY | Save Query | Save View | Format Query | Schedule Query | Show Options | Query complete (0.9s elapsed, cached)

Results | Details | Download as CSV | Download as JSON | Save as Table | Save to Google Sheets

Row	job_id	worksite_county	county	area	worksite_state	state	soc_code	soc_name	year	job_salary	national_salary
1	0000035e-b44f-4785-b2e4-a8d10e4e8857	TOMPKINS	TOMPKINS	27060	NY	NY	15-1131	COMPUTER PROGRAMMERS	2016	64000.0	109850.4
2	000013d8-82cd-4ba5-8d17-cbab48d8d5b6	KING	KING	42644	WA	WA	15-1121	COMPUTER SYSTEMS ANALYSTS	2017	84000.0	139138.0
3	00001e51-1306-46ff-8fed-e8173fb62d54	MARICOPA	MARICOPA	38060	AZ	AZ	15-1132	SOFTWARE DEVELOPERS, APPLICATIONS	2017	92435.0	132655.6
4	0000287a-c470-4130-858e-d428865b70ce	BENTON	BENTON	22220	AR	AR	15-1131	COMPUTER PROGRAMMERS	2018	65187.0	92534.8
5	000028aa-3eda-4314-afc9-d0281d114983	ERIE	ERIE	15380	NY	NY	11-3021	COMPUTER AND INFORMATION SYSTEMS MANAGERS	2016	105000.0	166761.2
6	00004038-4c15-4ede-b57c-e770c019ba63	LARIMER	LARIMER	22660	CO	CO	15-1132	SOFTWARE DEVELOPERS, APPLICATIONS	2018	77500.0	140452.0
7	0000403f-3bf5-4b69-a8cf-cd24a0bed4c4	COOK	COOK	16974	IL	IL	15-1134	WEB DEVELOPERS	2017	63419.0	110697.2
8	00005496-a906-41f0-a0f7-084eefa9628f	SAN FRANCISCO	SAN FRANCISCO	41884	CA	CA	15-1121	COMPUTER SYSTEMS ANALYSTS	2017	125000.0	167403.6

Table | JSON

First < Prev Rows 1 - 8 of 1200869 Next > Last

Cross-Dataset Views

v_Tech_Job:

- Filters out all non-tech jobs from the Job table

v_Tech_Job_Salary_Comparison:

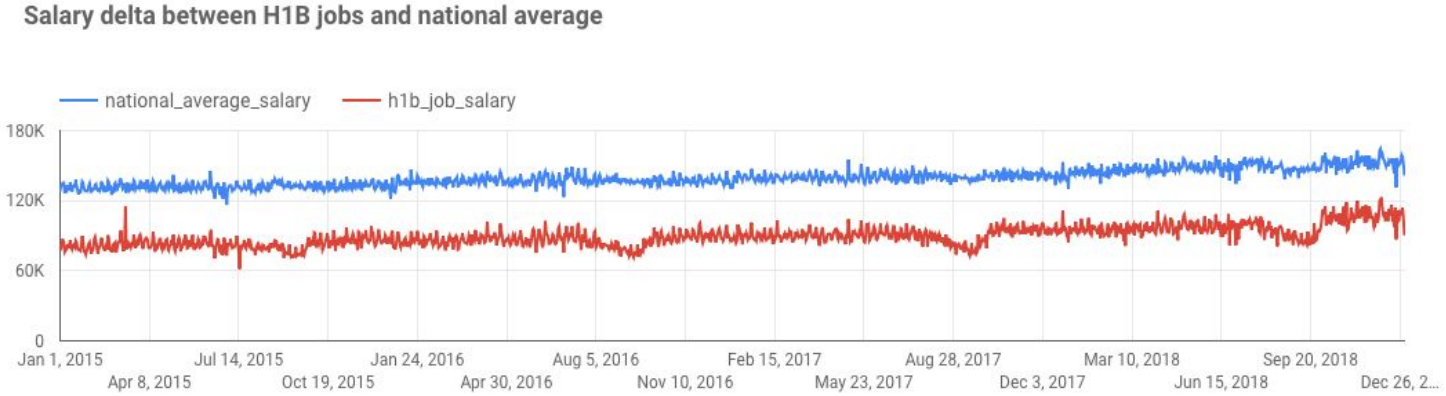
- 3-table join between Job, Geography, and Wages
- Calculates salary delta between h1b job and national average

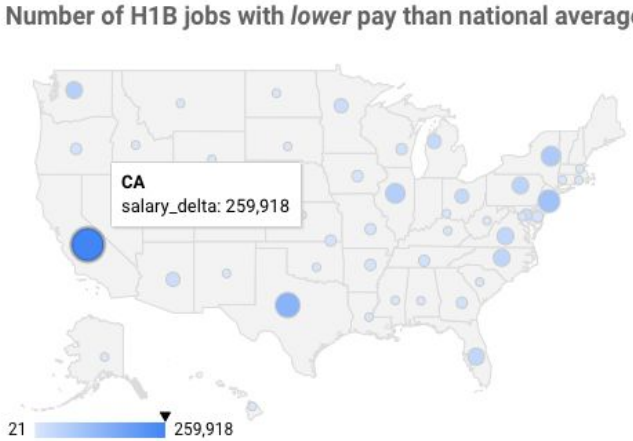
v_Tech_Job_Salary_Delta_by_Occupation:

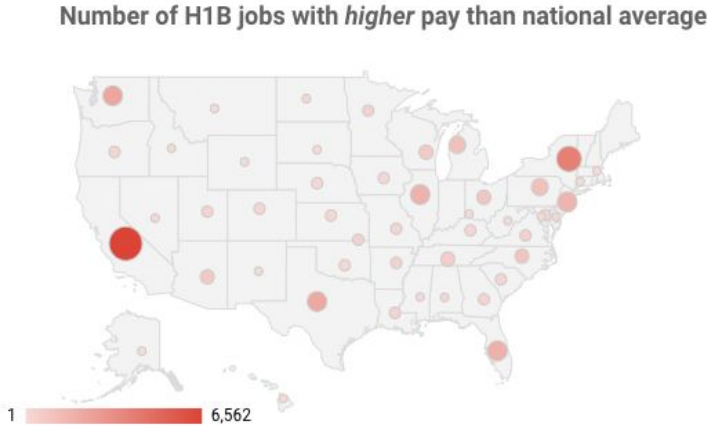
- Groups jobs by occupation
- Calculates salary delta as h1b job - national average

Source File: https://github.com/shirleycohen/h1b_analytics/blob/master/job_salary_views.sql

Data Studio Report



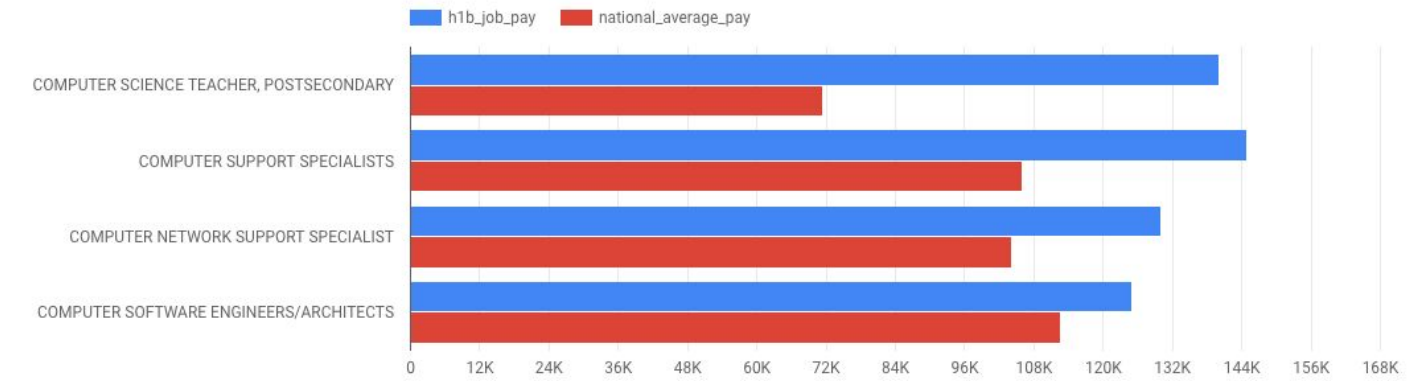




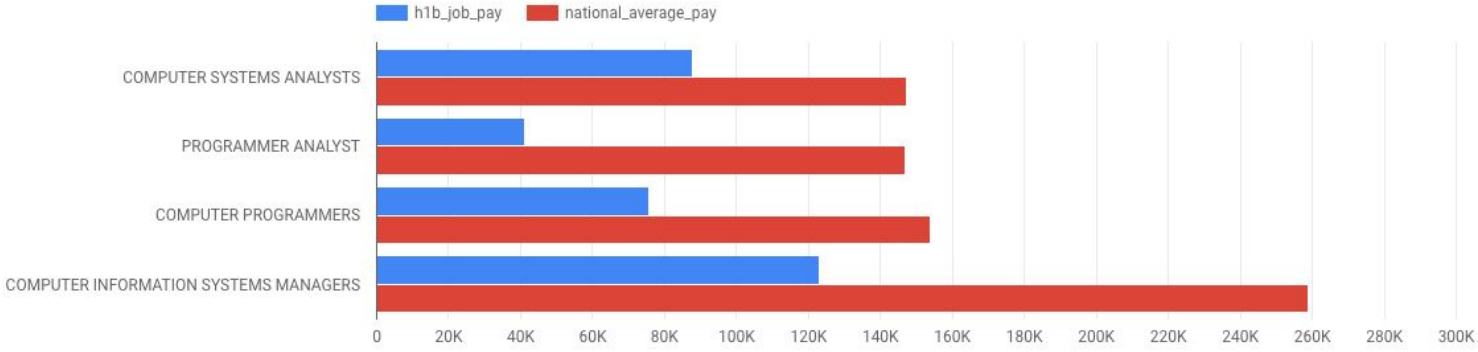
Data Studio Report

Pay Gaps by Occupation:

Occupations which pay H1B workers *higher* than domestic workers



Occupations which pay H1B workers *lower* than domestic workers



View Switching Technique

Using v_Tech_Job_Salary_Comparison as example:

- Create “shadow” tables from “main” Job, Geography, and Wages tables
- Populate “shadow” tables with new records (e.g. 2019) and changed records for time period 2015 - 2018
- Create “shadow” view that points to “shadow” tables
- Verify correctness of “shadow” tables by querying the “shadow” view
- Promote “shadow” tables to “main” tables
- Recreate “main” view to point to new “main” tables

Milestone 9

http://www.cs.utexas.edu/~scohen/milestones/Milestone9.pdf

Milestone 10

http://www.cs.utexas.edu/~scohen/milestones/Milestone10.pdf