# CS 327E Class 8

Oct 30, 2020

# Final Project Components

- Choose a primary and secondary dataset (Milestone 1)
- Load the raw data into BigQuery (Milestone 1)
- Explore the raw data with SQL (Milestone 1)
- Cleanse the data with SQL (Milestone 2)
- Create a unified model of the data (Milestone 2)
- Cleanse the data with Apache Beam (Milestone 3)
- Analyze the refined data with SQL (Milestone 4)
- Create data visualizations with Data Studio (Milestones 2, 3, 4)
- Present your work (Final Presentation)

**Primary Dataset:**
**H1B Visa applications**

Source:
US Dept. of Labor

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

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

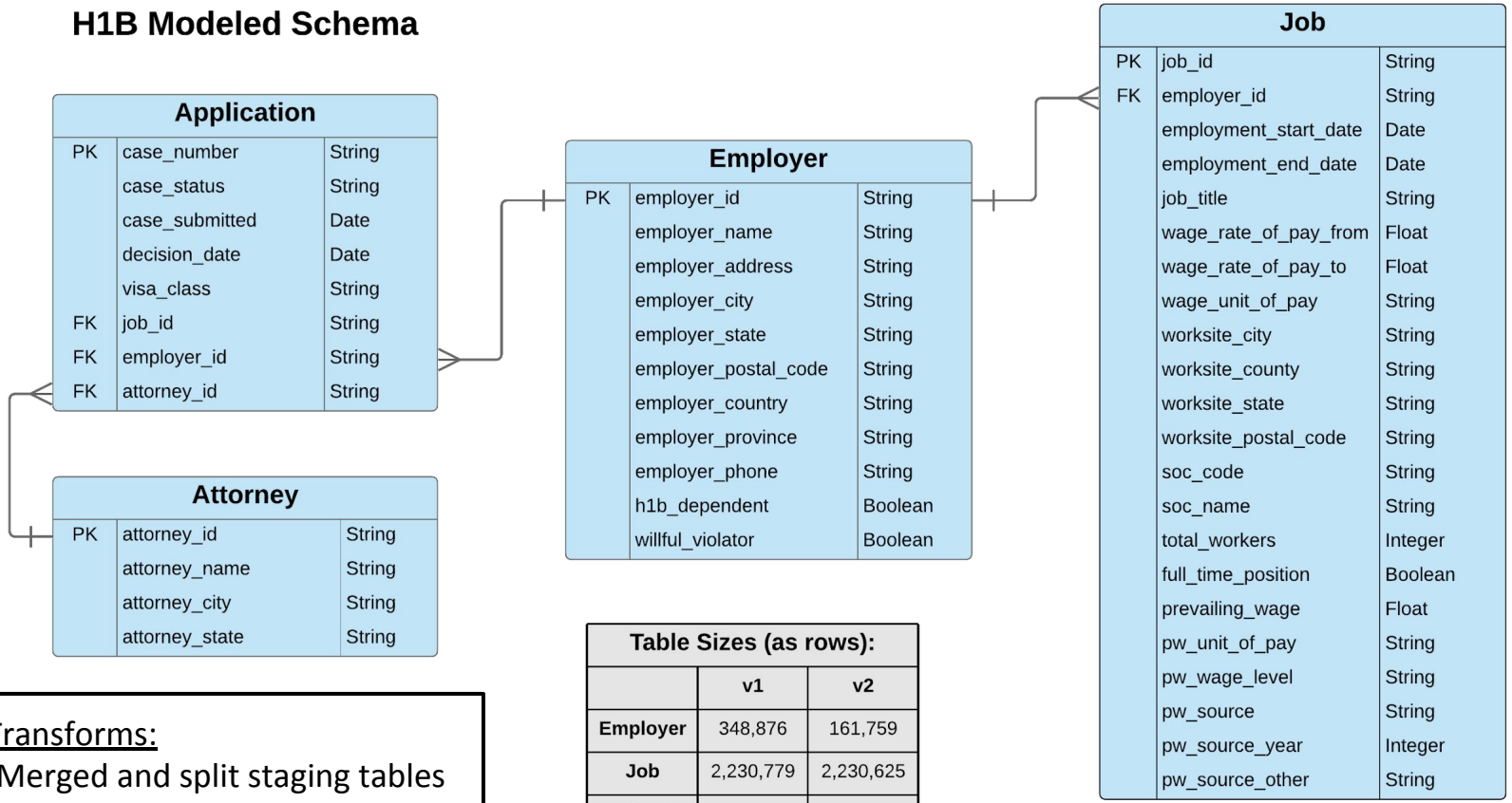
Project Work:
-Imported files into staging tables

**Table Details: H1B_Applications_2017**

Schema | Details | Preview

| | | |
|---|---|---|
| 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 Modeled Schema




| 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 |

Transforms:

-Merged and split staging tables

-Enforced referential integrity

-Removed duplicate records

**Secondary Dataset 1:**
**Corporate Registrations**

Source:
Secretary of State from 13 states

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

## Table Details: Corporate_Registrations_CA

Schema | Details | Preview

| | |
|---|---|
| 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 2:**
**Occupational Employment Survey**

Source: Bureau of Labor Statistics

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

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

Project Work:
-Imported files into staging tables

## 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 |

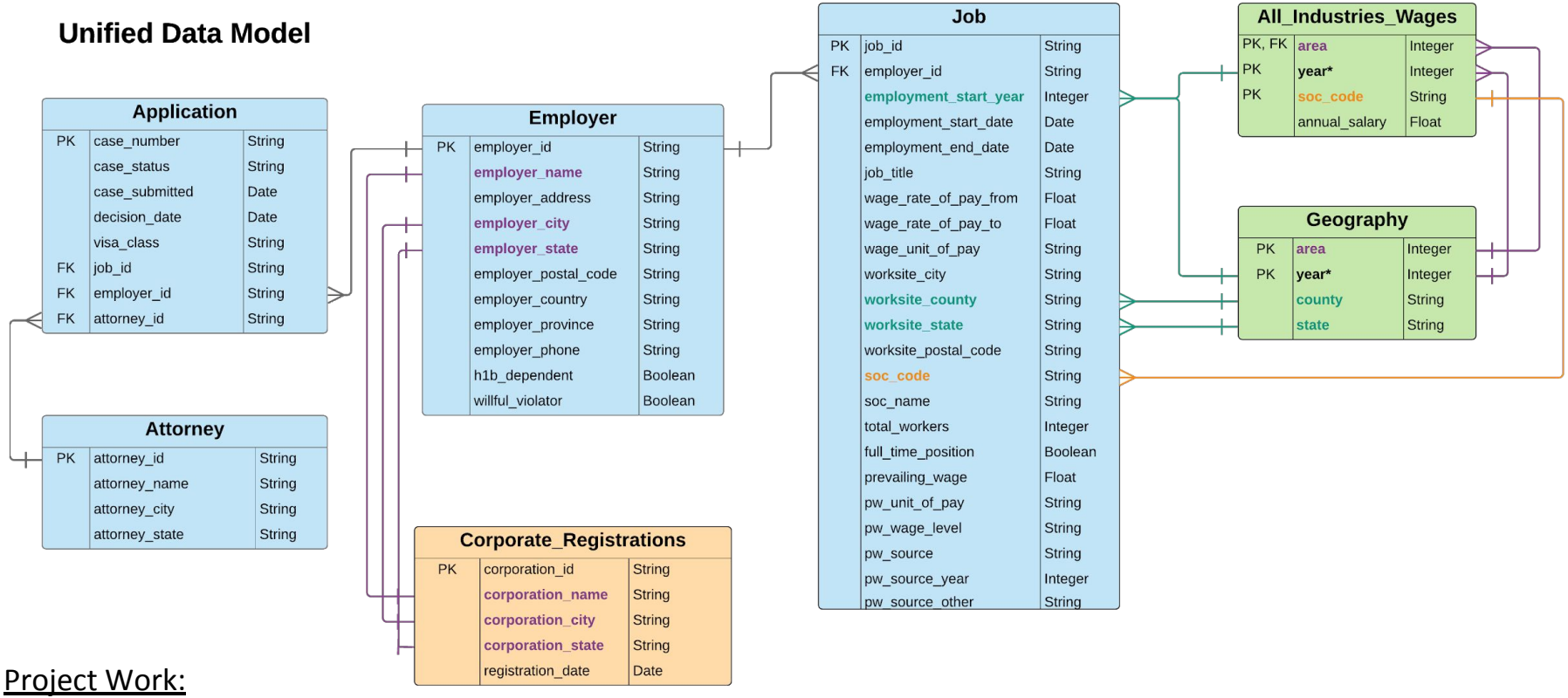
## Table Details: Geography_2018

Refresh | Query Table

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) |

# Unified Data Model




Project Work:

-Merged corp. registration tables

-Merged wages tables

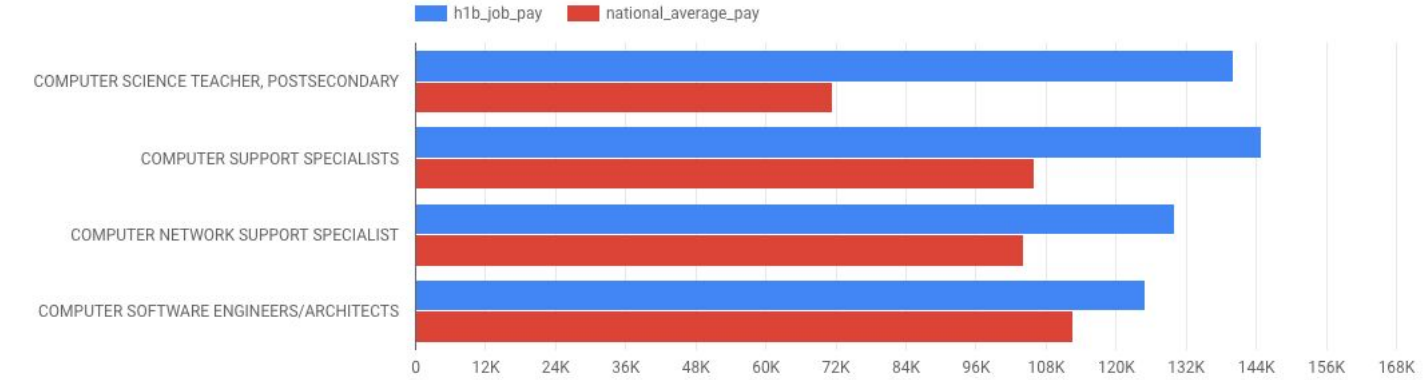
-Merged geography tables

-Normalized corporation name, city, state

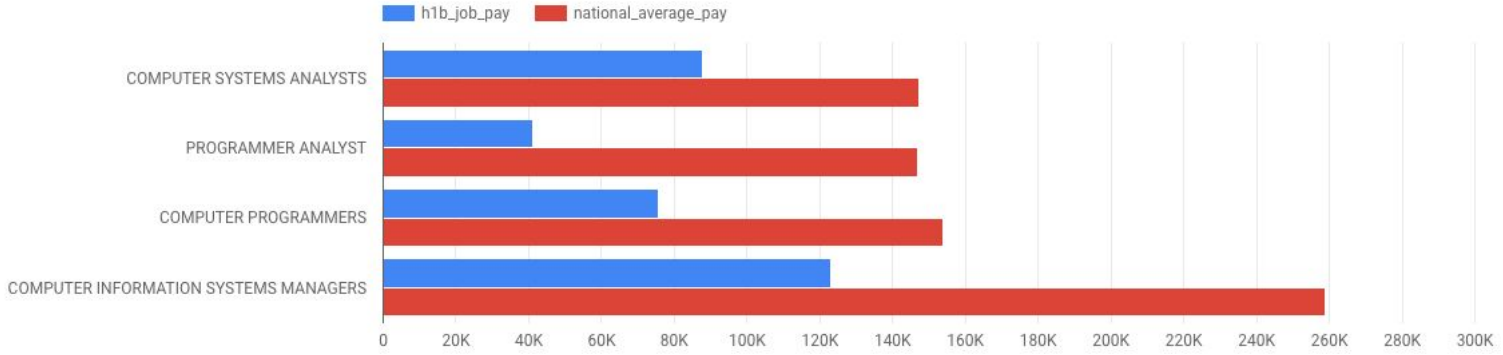
# Sample Reports

## Pay Gaps by Occupation:

### Occupations which pay H1B workers *higher* than domestic workers




### Occupations which pay H1B workers *lower* than domestic workers



# Dataset Listings

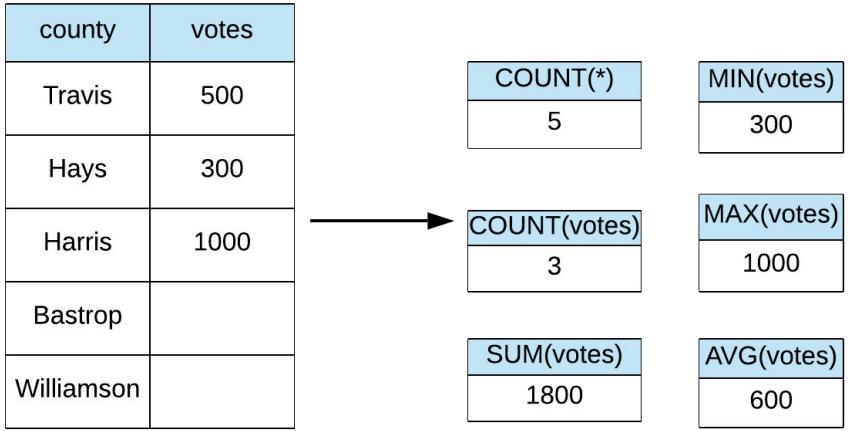
| Topic | Primary Dataset | Secondary Dataset |
|---|---|---|
| **Public Health** | COVID-19 cases (source: JHU daily reports) | American Community Survey (source: US Census Bureau) |
| **Transportation** | Airline on-time performance (source: Bureau of Transportation Statistics) | Storm events (source: NOAA) |
| **Housing** | Short-term rentals in 30+ cities (source: Airbnb) | Long-term rentals nationwide (source: Zillow) |
| **Employment** | H1B visa applications (source: US Department of Labor) | Business registrations (source: Secretary of State for various states)<br>Occupational Employment Survey (source: Bureau of Labor Statistics) |
| **Movies** | Hollywood Movies, Directors, Actors (source: IMDB) | Bollywood Movies, Actors and Songs (source: Kaggle) |
| **Music** | Artists and Songs (source: MusicBrainz) | Artists, Labels, Recordings (source: Discog) |

# Global Aggregate Queries

```
SELECT <aggregate function>
       [, <aggregate function>]
FROM <single table>
[JOIN <single table>
 ON <join condition>]
[WHERE <boolean condition>]
ORDER BY <field(s) to sort on>
```

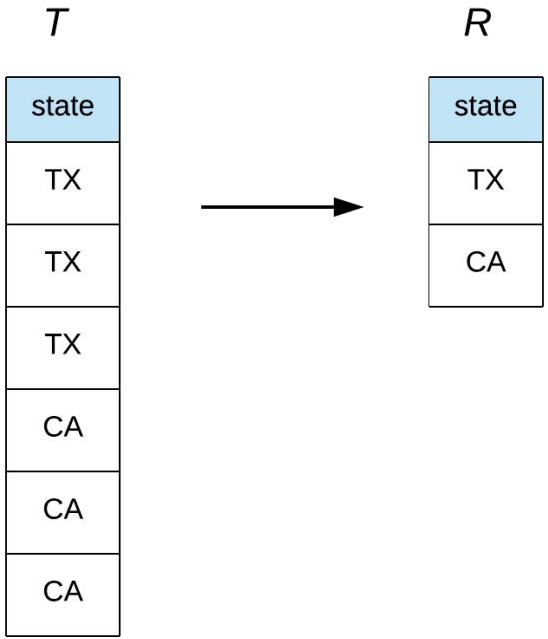
# Global Aggregate Queries

```
SELECT <aggregate function>
       [, <aggregate function>]
FROM <single table>
[JOIN <single table>
 ON <join condition>]
[WHERE <boolean condition>]
ORDER BY <field(s) to sort on>
```



# Group By Queries

```
SELECT <unaggregated field(s)>
FROM <single table>
[JOIN <single table>
ON <join condition>]
[WHERE <boolean condition>]
GROUP BY <unaggregated field(s)>
```

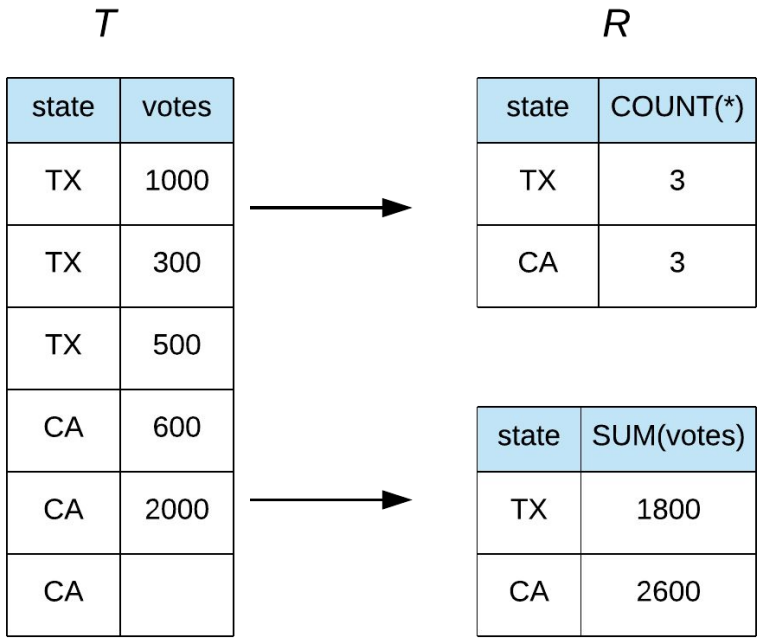


# Aggregate Group By Queries

```
SELECT <unaggregated field(s)>,
       <aggregate function(s)>
FROM <single table>
[JOIN <single table>
 ON <join condition>]
[WHERE <boolean condition>]
GROUP BY <unaggregated field(s)>
[HAVING <boolean condition>]
[ORDER BY <field(s) to sort on>]
```

# Aggregate Group By Queries

```
SELECT <unaggregated field(s)>,
       <aggregate function(s)>
FROM <single table>
[JOIN <single table>
 ON <join condition>]
[WHERE <boolean condition>]
GROUP BY <unaggregated field(s)>
[HAVING <boolean condition>]
[ORDER BY <field(s) to sort on>]
```



# How to COUNT

```
SELECT COUNT(*)
FROM Employee

SELECT COUNT(department)
FROM Employee

SELECT DISTINCT department
FROM Employee

SELECT COUNT(DISTINCT department)
FROM Employee
```

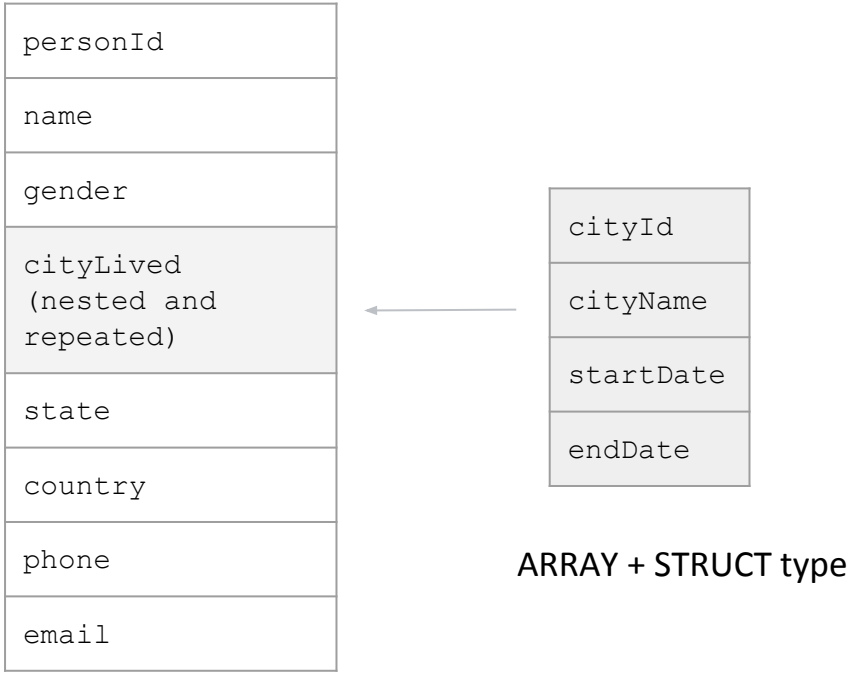
**Employee**

| row | employee | department |
|---|---|---|
| 1 | Sunil | ENG |
| 2 | Morgan | ENG |
| 3 | Rama | Product |
| 4 | Drew | |
| 5 | Jeff | Research |
| 6 | Danielle | HR |
| 7 | Grace | ENG |

# Why BigQuery?

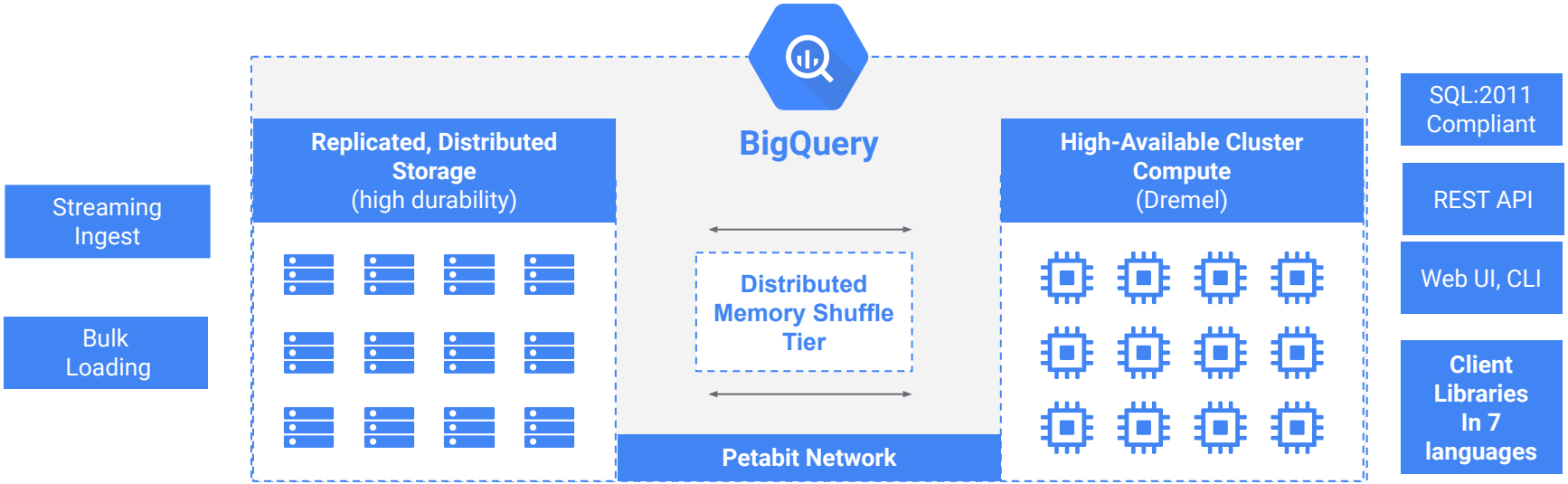
- Analytics database service on GCP
- Designed for storing and querying large data (petabyte-scale)
- Tables stored in columnar layout
- ANSI SQL compliant
- Data Types:
  - Primitive: BOOL, BYTES, FLOAT64, INT64, NUMERIC, STRING
  - Temporal: DATE, DATETIME, TIME, TIMESTAMP
  - Geospatial: GEOGRAPHY
  - Complex:  ARRAY, STRUCT
- Not designed for transaction-heavy workloads
- No built-in referential integrity

# Nested Columns



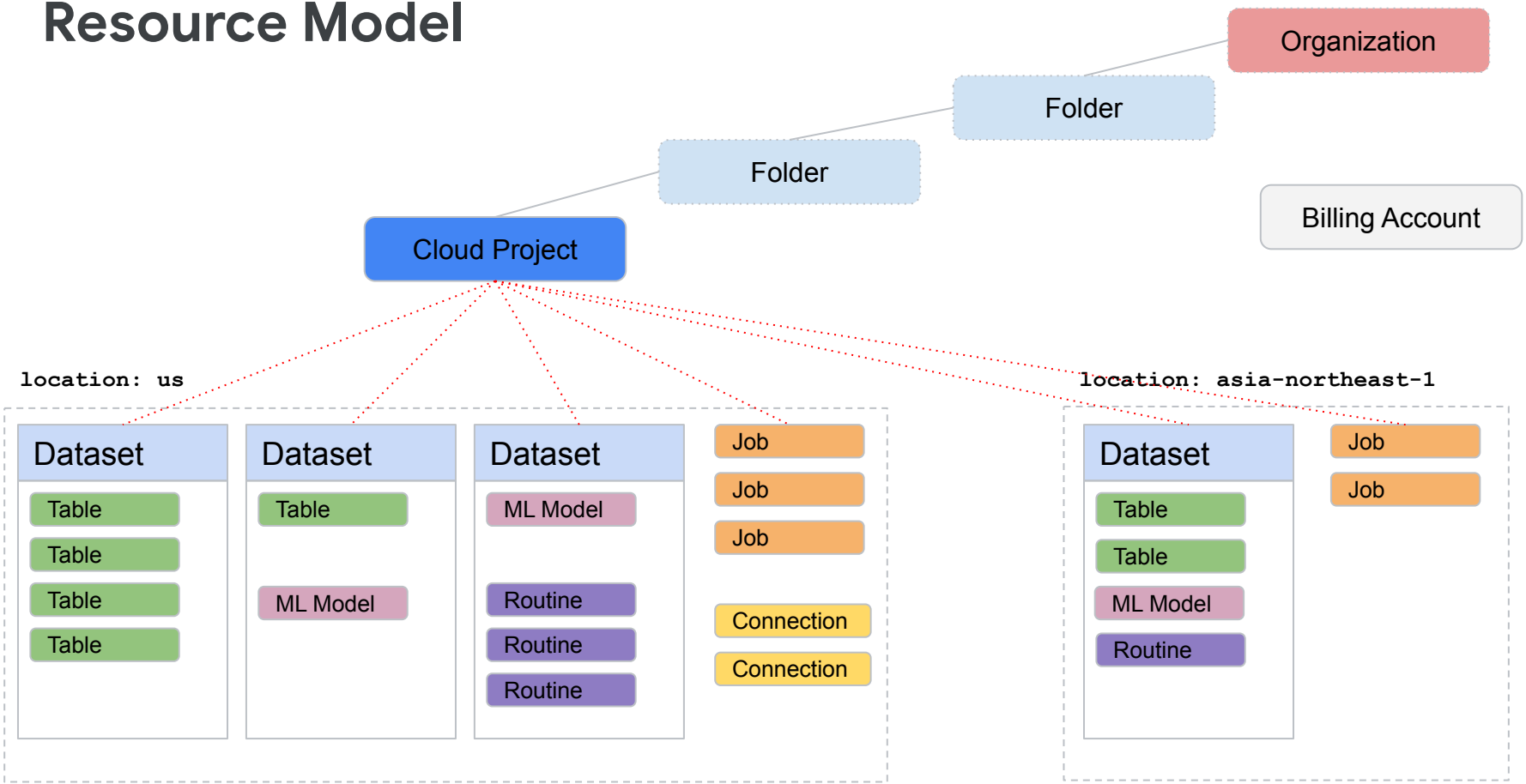

ARRAY + STRUCT type

# High-level Architecture




Google Cloud

# Resource Model



# Getting Started with BigQuery

No setup guide needed :)

# Practice Problems

1. For each class, how many students are enrolled in the class?

   Return the cno and count for each class.
2. For each class which has at least two students enrolled, how many students are taking the class?

Student(sid, fname, lname, dob, status)

Class(cno, cname, credits)

Teacher(tid, fname, lname, dept)

Takes(sid, cno, grade)

Teaches(tid, cno)

# Milestone 1

http://www.cs.utexas.edu/~scohen/projects/Milestone1.pdf