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

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Nullable</th>
</tr>
</thead>
<tbody>
<tr>
<td>case_number</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>visa_class</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>case_status</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_name</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_business_db</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_address</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_city</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_state</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_postal_code</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_country</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_province</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_phone</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employer_phone_ext</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>naics_code</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>soc_name</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>soc_code</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>job_title</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>total_workers</td>
<td>INTEGER</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>case_submitted</td>
<td>TIMESTAMP</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>decision_date</td>
<td>TIMESTAMP</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employment_start_date</td>
<td>TIMESTAMP</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>employment_end_date</td>
<td>TIMESTAMP</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>full_time_position</td>
<td>BOOLEAN</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>prevailing_wage</td>
<td>FLOAT</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>pw_unit_of_pay</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>wage_rate_of_pay_from</td>
<td>FLOAT</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>wage_rate_of_pay_to</td>
<td>FLOAT</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>wage_unit_of_pay</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>worksite_city</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>worksite_county</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>worksite_state</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>worksite_postal_code</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>agent_attorney_name</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>agent_attorney_city</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>agent_attorney_state</td>
<td>STRING</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>h1b_dependent</td>
<td>BOOLEAN</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>willful_violator</td>
<td>BOOLEAN</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>original_cert_date</td>
<td>TIMESTAMP</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>new_employment</td>
<td>FLOAT</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>continued_employment</td>
<td>FLOAT</td>
<td>NULLABLE</td>
</tr>
<tr>
<td>change_previous_employment</td>
<td>FLOAT</td>
<td>NULLABLE</td>
</tr>
</tbody>
</table>
### H1B Modeled Schema

#### Application
- **PK**: case_number, case_status, case_submitted, decision_date, visa_class, job_id, employer_id, attorney_id
- **FK**: job_id, employer_id, attorney_id
- **Attributes**:
  - String: case_number, case_status, case_submitted, decision_date, visa_class, employer_id, attorney_id
  - Date: decision_date
  - Boolean: willful_violator

#### Employer
- **PK**: employer_id
- **Attributes**:
  - String: employer_id, employer_name, employer_address, employer_city, employer_state, employer_country, employer_province, employer_phone, h1b_dependent, willful_violator

#### Job
- **PK**: job_id
- **Attributes**:
  - String: job_id, employer_id
  - Date: employment_start_date, employment_end_date, job_title
  - Float: wage_rate_of_pay_from, wage_rate_of_pay_to, wage_unit_of_pay
  - String: worksite_city, worksite_state, worksite_county, worksite_postal_code

#### Attorney
- **PK**: attorney_id
- **Attributes**:
  - String: attorney_id, attorney_name, attorney_city, attorney_state

### Table Sizes (as rows):

<table>
<thead>
<tr>
<th>Table</th>
<th>v1</th>
<th>v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employer</td>
<td>348,876</td>
<td>161,759</td>
</tr>
<tr>
<td>Job</td>
<td>2,230,779</td>
<td>2,230,625</td>
</tr>
<tr>
<td>Application</td>
<td>2,633,426</td>
<td>2,633,156</td>
</tr>
<tr>
<td>Attorney</td>
<td>19,861</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>so_file_number</td>
<td>STRING</td>
</tr>
<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_county</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>termExpiration_date</td>
<td>DATE</td>
</tr>
<tr>
<td>chiefExecutiveOfficerName</td>
<td>STRING</td>
</tr>
<tr>
<td>chiefExecutiveOfficerAddressLine_1</td>
<td>STRING</td>
</tr>
<tr>
<td>chiefExecutiveOfficerAddressLine_2</td>
<td>STRING</td>
</tr>
<tr>
<td>chiefExecutiveOfficerAddressCity</td>
<td>STRING</td>
</tr>
<tr>
<td>chiefExecutiveOfficerAddressStateOrCounty</td>
<td>STRING</td>
</tr>
<tr>
<td>chiefExecutiveOfficerAddressZipCode</td>
<td>STRING</td>
</tr>
<tr>
<td>agentName</td>
<td>STRING</td>
</tr>
<tr>
<td>agentAddressLine_1</td>
<td>STRING</td>
</tr>
<tr>
<td>agentAddressLine_2</td>
<td>STRING</td>
</tr>
<tr>
<td>agentAddressCity</td>
<td>STRING</td>
</tr>
<tr>
<td>agentAddressStateOrCounty</td>
<td>STRING</td>
</tr>
<tr>
<td>agentAddressZipCode</td>
<td>STRING</td>
</tr>
<tr>
<td>stateOrForeignCountry</td>
<td>STRING</td>
</tr>
<tr>
<td>ftbSuspensionStatus</td>
<td>STRING</td>
</tr>
<tr>
<td>corporationTaxBase</td>
<td>STRING</td>
</tr>
<tr>
<td>transactionJulianDate</td>
<td>DATE</td>
</tr>
<tr>
<td>ftbSuspensionString</td>
<td>STRING</td>
</tr>
<tr>
<td>filler</td>
<td>STRING</td>
</tr>
</tbody>
</table>
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
Project Work:
- Merged corp. registration tables
- Merged wages tables
- Merged geography tables
- Normalized corporation 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
## Dataset Listings

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

$$\text{SELECT} \ <\text{unaggregated field(s)}>, \ <\text{aggregate function(s)}>$$

$$\text{FROM} \ <\text{single table}>$$

$$[\text{JOIN} \ <\text{single table}> \ \text{ON} \ <\text{join condition}>]$$

$$[\text{WHERE} \ <\text{boolean condition}>]$$

$$\text{GROUP BY} \ <\text{unaggregated field(s)}>$$

$$[\text{HAVING} \ <\text{boolean condition}>]$$

$$[\text{ORDER BY} \ <\text{field(s) to sort on}>]$$

\[
\begin{array}{|c|c|}
\hline
\text{state} & \text{votes} \\
\hline
\text{TX} & 1000 \\
\text{TX} & 300 \\
\text{TX} & 500 \\
\text{CA} & 600 \\
\text{CA} & 2000 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{state} & \text{COUNT}(*) \\
\hline
\text{TX} & 3 \\
\text{CA} & 3 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{state} & \text{SUM(votes)} \\
\hline
\text{TX} & 1800 \\
\text{CA} & 2600 \\
\hline
\end{array}
\]
How to COUNT

SELECT COUNT(*)
FROM Employee

SELECT COUNT(department)
FROM Employee

SELECT DISTINCT department
FROM Employee

SELECT COUNT(DISTINCT department)
FROM Employee
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

<table>
<thead>
<tr>
<th>personId</th>
<th>name</th>
<th>gender</th>
<th>cityLived (nested and repeated)</th>
<th>state</th>
<th>country</th>
<th>phone</th>
<th>email</th>
</tr>
</thead>
</table>

**ARRAY + STRUCT type**

- cityId
- cityName
- startDate
- endDate
High-level Architecture

Streaming Ingest

Bulk Loading

Replicated, Distributed Storage (high durability)

Distributed Memory Shuffle Tier

Petabit Network

BigQuery

High-Available Cluster Compute (Dremel)

SQL:2011 Compliant

REST API

Web UI, CLI

Client Libraries In 7 languages

Google Cloud
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