

# H-1B VISA

Deriving insights from data in order to get better understanding of H-1B VISA.

YKDBProject (Hyeon Gu Kim, Junho Yang)





## Area of Interest

- Which areas of occupation, or occupation itself, have frequently requested for H-1B?
- Which ownership (private or local government, etc..) of occupation tends to request for H-1B petition more?
- Derive informative insight that may be helpful to international students like ourselves.



# Datasets

(From 2014 to 2018)

- **H-1B VISA Application dataset**

(source: US Department of Labor)

- **Occupational Employment Statistics datasets** (source: US Bureau of Labor Statistics)
  - Ownership, National, and Sector
- **USCIS Staging dataset**

(source: US Citizenship and Immigration Services)

# Milestone 1

- Created bucket in Google Cloud Storage and uploaded datasets in CSV format
- Imported the datasets into Google Big Query through Jupyter Notebook
- SQL Queries : Explored the datasets and made sure everything is working fine





## Milestone 2



Google  
BigQuery

- Created a new Big Query dataset to store our modeled tables
  - Split/join ...
  - H-1B Employer dataset - employer name, employer city,...employer country
  - H-1B Application dataset - case number, case status,...,case number, case status, case submitted
  - H-1B Occupation dataset - soc\_code, soc\_title,..., prevailing\_wage\_YR
- Chose a primitive data type that most precisely represents
  - Basic data cleansing for using Apache Beam in the next milestone.
  - changed yyyy.mm.dd STRING format to yyyy-mm-dd format so that we can CAST its type to DATE.  
(CASE\_SUBMITTED and DECESION\_DATE)
- Created views for data visualization in Dataflow later.



## Milestone 3



- Apache Beam makes it easy to describe the various aspects of the out-of-order processing
- Identify all the tables from our refined datasets which contain data that needs to be cleansed.
  - “employment\_start\_date” & “employment\_end\_date”
    - transform 2012.12.03 & 12/03/2012 to **2012-12-03** format
  - “soc\_code”
    - transform 12-3456.00, 30, 12.3456.00 to **12-3456** format
  - other tables:
    - remove duplicates
- Write a Beam pipeline that normalizes the data from the table
- PK/FK Verification process

## Milestone 4

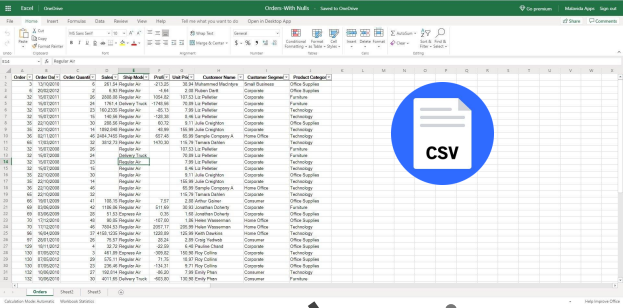
- Converted Beam pipelines to Dataflow
- Verified each tables have valid PK or FK
- Updated ERD
- 3 Cross-dataset queries
- Created data visualization in Data Studio



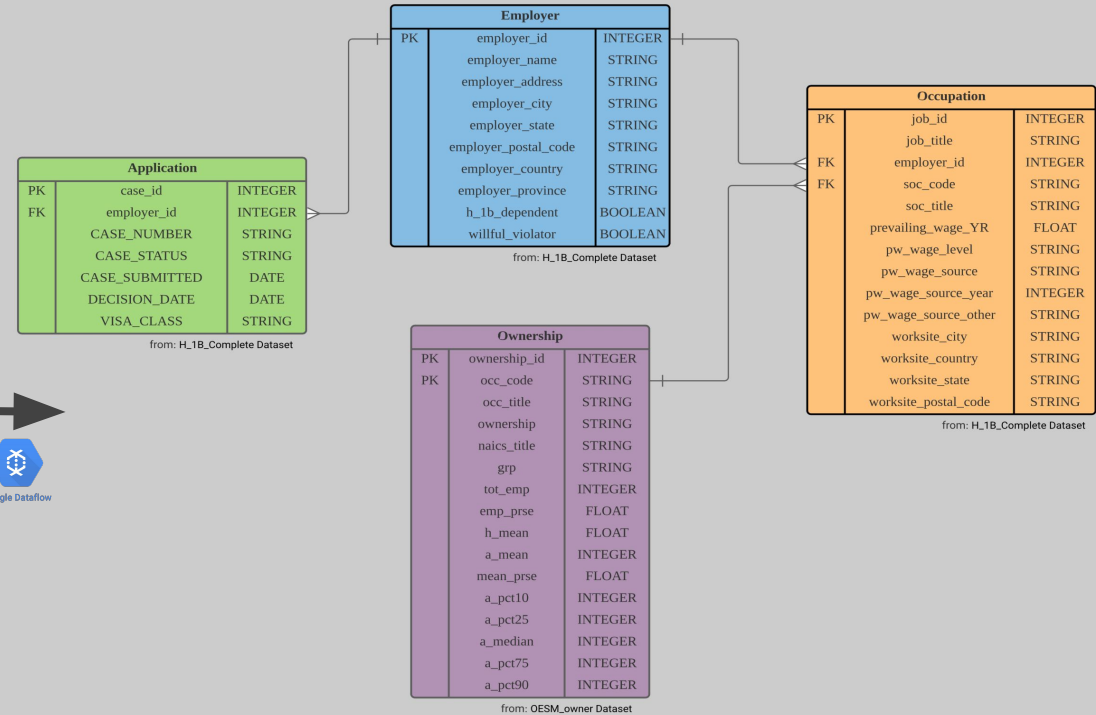
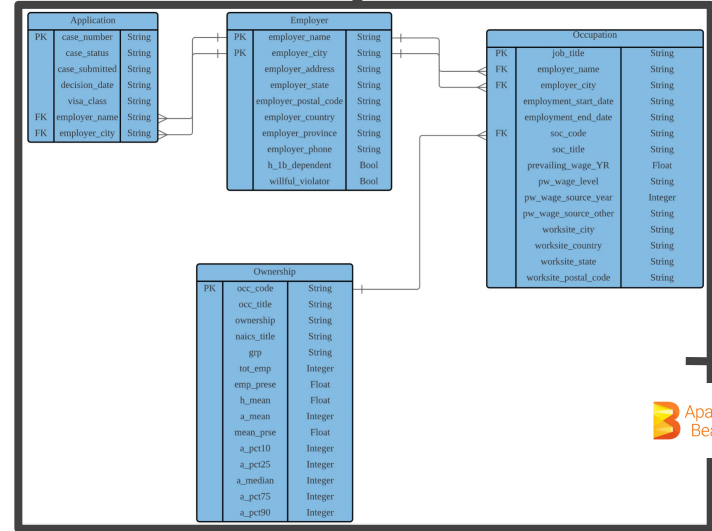
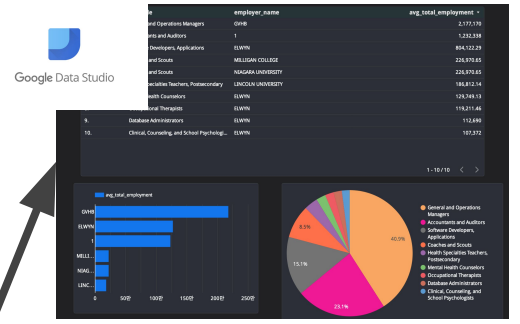
Google Dataflow



Google Data Studio



# Overview







# Future Improvements

- Income difference of those who got certified and those who are denied.
- Certified and Denied proportion over the years
- Optimize SQL queries (especially JOIN clauses)





**Thank you!**