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**

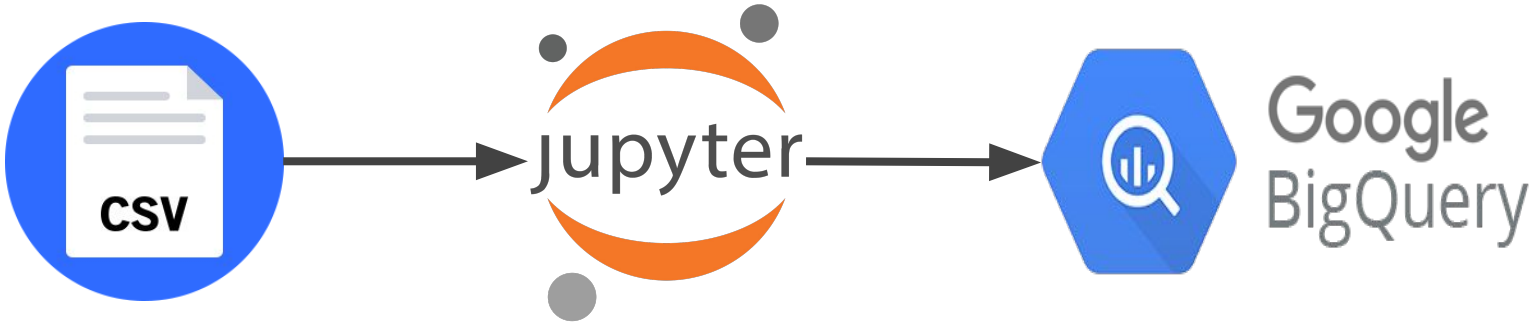
  (source: US Department of Labor)

- **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

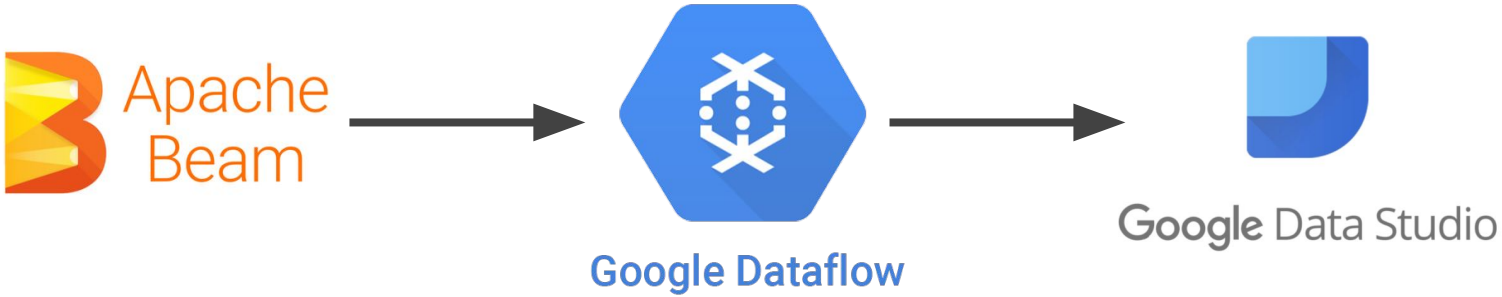
- Created a new Big Query dataset to store our modeled tables
  - Split/join ...
  - H-1B Employer dataset - employer name, employer city,..,employer country
  - H-1B Application dataset - case number, case status,..,case number, case status, case submitted
  - H-1B Occupation dataset - soc_code, soc_title,..., 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

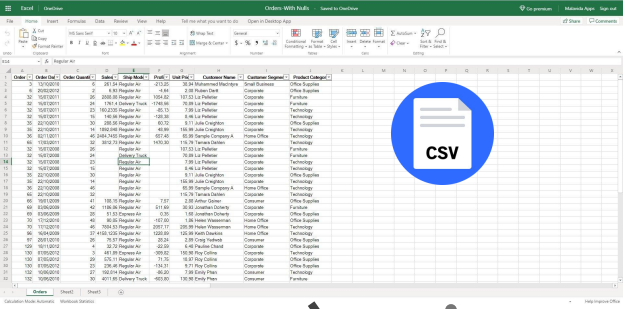
- Apache Beam makes it easy to describe the various aspects of the out-of-order processing
- Identify all the tables from our refined datasets which contain data that needs to be cleansed.
  - “employment_start_date” & “employment_end_date”
    - transform 2012.12.03 & 12/03/2012 to **2012-12-03** format
  - **“**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



Overview



Google Data Studio

jupyter

Google BigQuery

Application

PK	case_number	String
	case_status	String
	case_submitted	String
	decision_date	String
	visa_class	String
FK	employer_name	String
FK	employer_city	String

Employer

PK	employer_name	String
PK	employer_city	String
	employer_address	String
	employer_state	String
	employer_postal_code	String
	employer_country	String
	employer_province	String
	employer_phone	String
	h_1b_dependent	Bool
	willful_violator	Bool

Occupation

PK	job_title	String
FK	employer_name	String
FK	employer_city	String
	employment_start_date	String
	employment_end_date	String
FK	soc_code	String
	soc_title	String
	prevailing_wage_YR	Float
	pw_wage_level	String
	pw_wage_source_year	Integer
	pw_wage_source_other	String
	worksite_city	String
	worksite_country	String
	worksite_state	String
	worksite_postal_code	String

Ownership

PK	occ_code	String
	occ_title	String
	ownership	String
	naics_title	String
	grp	String
	tot_emp	Integer
	emp_prse	Float
	h_mean	Float
	a_mean	Integer
	mean_prse	Float
	a_pct10	Integer
	a_pct25	Integer
	a_median	Integer
	a_pct75	Integer
	a_pct90	Integer

Apache Beam

Google Dataflow

Application

PK	case_id	INTEGER
FK	employer_id	INTEGER
	CASE_NUMBER	STRING
	CASE_STATUS	STRING
	CASE_SUBMITTED	DATE
	DECISION_DATE	DATE
	VISA_CLASS	STRING

from: H_1B_Complete Dataset

Employer

PK	employer_id	INTEGER
	employer_name	STRING
	employer_address	STRING
	employer_city	STRING
	employer_state	STRING
	employer_postal_code	STRING
	employer_country	STRING
	employer_province	STRING
	h_1b_dependent	BOOLEAN
	willful_violator	BOOLEAN

from: H_1B_Complete Dataset

Occupation

PK	job_id	INTEGER
	job_title	STRING
FK	employer_id	INTEGER
FK	soc_code	STRING
	soc_title	STRING
	prevailing_wage_YR	FLOAT
	pw_wage_level	STRING
	pw_wage_source	STRING
	pw_wage_source_year	INTEGER
	pw_wage_source_other	STRING
	worksite_city	STRING
	worksite_country	STRING
	worksite_state	STRING
	worksite_postal_code	STRING

from: H_1B_Complete Dataset

Ownership

PK	ownership_id	INTEGER
PK	occ_code	STRING
	occ_title	STRING
	ownership	STRING
	naics_title	STRING
	grp	STRING
	tot_emp	INTEGER
	emp_prse	FLOAT
	h_mean	FLOAT
	a_mean	INTEGER
	mean_prse	FLOAT
	a_pct10	INTEGER
	a_pct25	INTEGER
	a_median	INTEGER
	a_pct75	INTEGER
	a_pct90	INTEGER

from: OESM_owner Dataset

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!