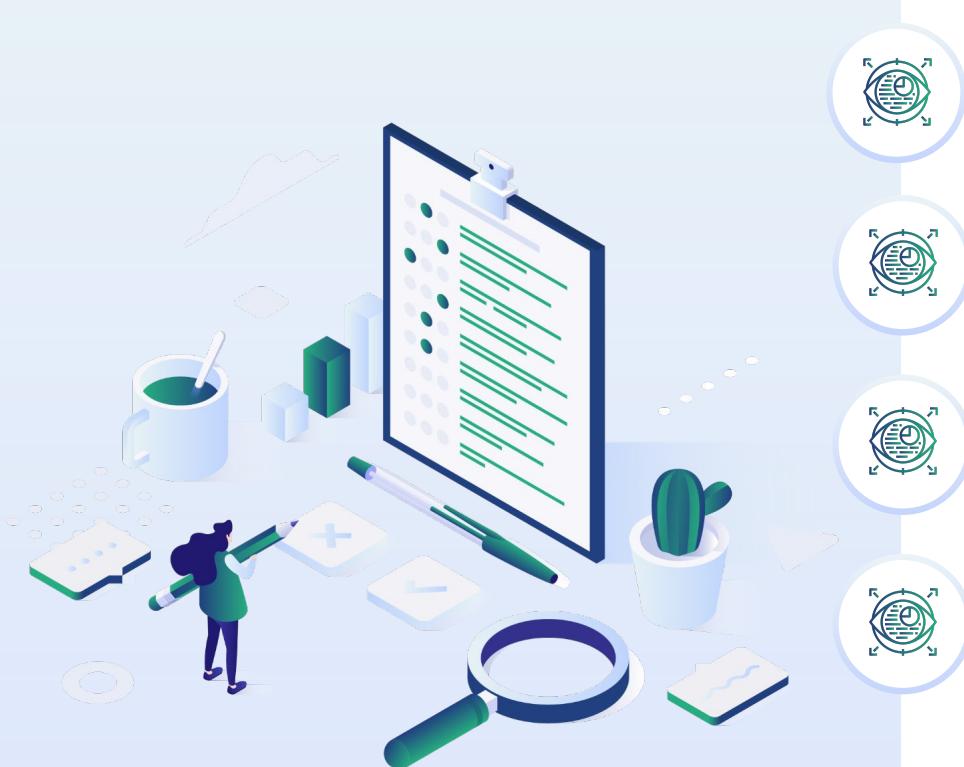


# EMPLOYEE ONSITE OPPORTUNITY





PROJECT OVERVIEW

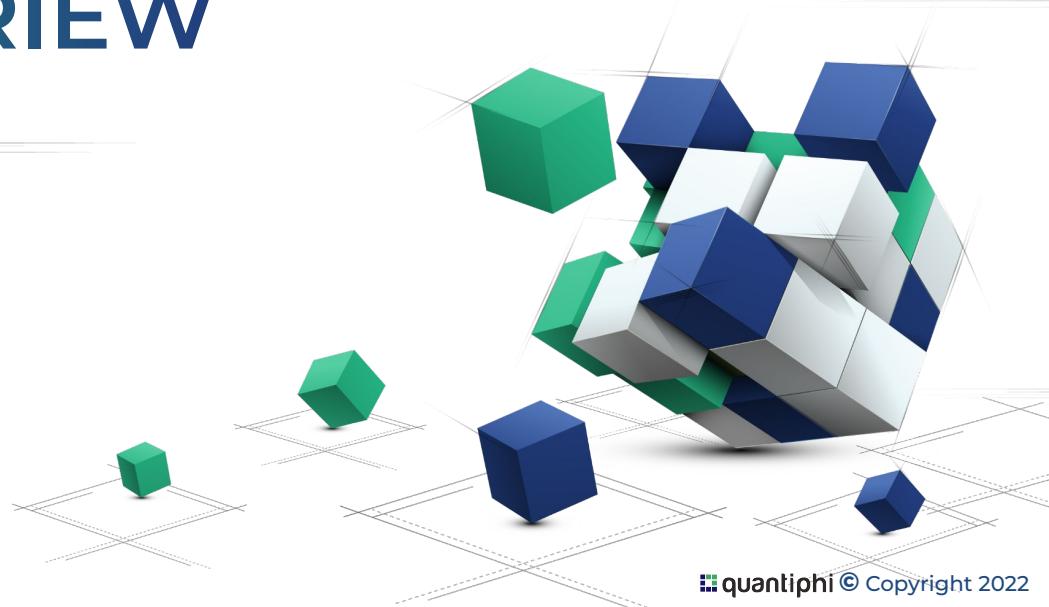
WORKFLOW

IMPLEMENTATION

CONCLUSION

# 01

## PROJECT OVERVIEW



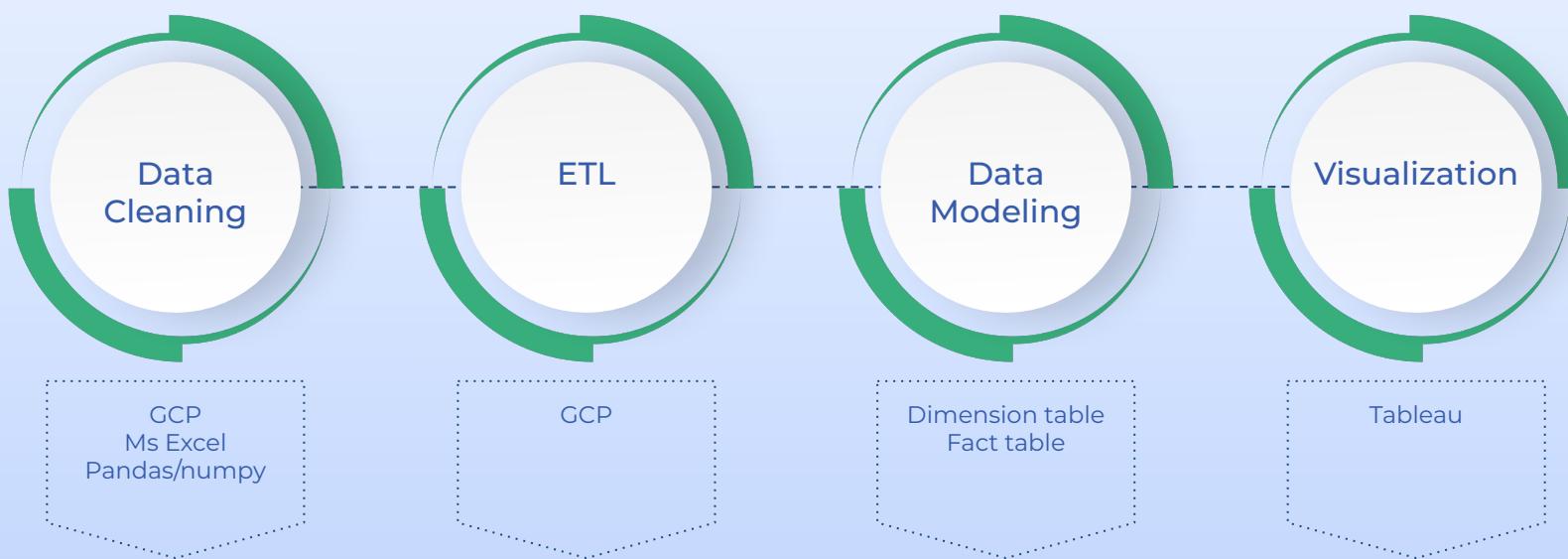
- ❖ We have some data on employment onsite opportunity
- ❖ The data describes about non immigrant visa that allows U.S companies and organizations to temporarily employ foreign workers in certain occupations.
- ❖ Data includes information on employers who have applied for employment onsite opportunity.

# 02

## WORK FLOW



# WORK FLOW



# 03

## IMPLEMENTATION



- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled
- Tools used:
  - Ms excel
  - Python pandas and numpy
  - GCP Data Prep

Google Cloud My First Project Search for resources, docs, products, and more (/) Search

STOP IMPORT AS PIPELINE SHARE

Dataflow

Overview Jobs Pipelines Workbench Snapshots SQL Workspace

cloud-dataprep-untitled-flow-2-16732560-by-afzalullarce

JOB GRAPH EXECUTION DETAILS JOB METRICS RECOMMENDATIONS

Job steps view Graph view CLEAR SELECTION

PFilterTransform15 Succeeded 17 sec 1 of 1 stage succeeded

PMapTransform4 Succeeded 30 sec 1 of 1 stage succeeded

PAggregateTransform Succeeded 2 sec 3 of 3 stages succeeded

PTableStore...nsformGCS2~ Succeeded 1 sec 5 of 5 stages succeeded

PMapTransform3 Succeeded 0 sec 1 of 1 stage succeeded

PTableStore...nsformGCS~ Succeeded 8 sec 4 of 4 stages succeeded

PMapTransform5 Succeeded 27 sec 1 of 1 stage succeeded

PMapTransform6 Succeeded 2 sec 1 of 1 stage succeeded

PProfileTransform Succeeded 40 sec 7 of 7 stages succeeded

PTableStore...nsformGCS3~ Succeeded 1 sec 5 of 5 stages succeeded

**Job info**

Job name cloud-dataprep-untitled-flow-2-16732560-by-afzalullarce  
Job ID 2022-11-29\_21\_03\_27-1526367896777640639  
Job type Batch  
Job status Succeeded  
SDK version Apache Beam SDK for Java 2.35.0  
ⓘ A never version of the SDK family exists and updating is recommended.  
[Learn more](#)

Job region us-central1  
Worker location us-central1  
Current workers 0  
Latest worker status Worker pool stopped.  
Start time November 30, 2022 at 10:33:30 AM GMT+5  
Elapsed time 7 min 35 sec  
Encryption type Google-managed key  
Dataflow Prime Disabled  
Runner v2? Disabled  
Dataflow Shuffle Enabled

**Resource metrics**

Current vCPUs 1  
Total vCPU time 0.099 vCPU hr  
Current memory 3.75 GB  
Total memory time 0.371 GB hr  
Current HDD PD 250 GB  
Total HDD PD time 24.717 GB hr  
Current SSD PD 0 B  
Total SSD PD time 0 GB hr  
Total Shuffle data processed 38.01 MB  
Billable Shuffle data processed 9.5 MB

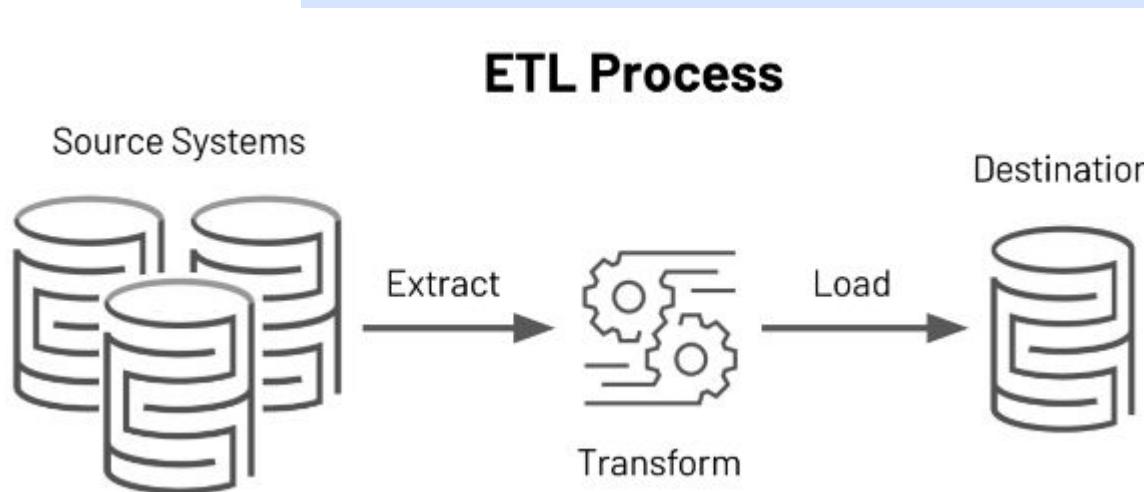
**Custom counters**

Filter Filter by counter name, value or step

Release Notes

quantiphi © Add copyright statement 2022 here.

ETL, which stands for extract, transform, and load, is the process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.



- 1) Create a bucket by CLI
  - a) gcloud storage buckets create gs://mock\_pro\_bucket1
- 2) Import cleaned CSV file to the bucket
- 3) Extract this into Wrangler
- 4) Wrangler - Transform
  - a) Build a robust batch data pipeline using data fusion with wrangler
- 5) Load this transformed data into BigQuery( Data warehouse)



# Data warehouse- BigQuery

Google Cloud My First Project Search Products, resources, docs (/)

Explorer + ADD DATA mock\_pro\_tb

Type to search

Viewing all resources. Show starred resources only.

SCHEMA DETAILS PREVIEW

Row	case_number	phase	render_date	sponsorship_type	appointment_start_date	appointment_end_date
1	I-200-09334-520385	CERTIFIED	11/30/2009	H-1B	12-11-2009	12-10-2012
2	I-200-10223-221072	WITHDRAWN	08-11-2010	H-1B	09-01-2010	6/30/2011
3	I-200-10228-922855	CERTIFIED	8/16/2010	H-1B	09-01-2010	6/30/2011
4	I-200-09315-469049	CERTIFIED	11/20/2009	H-1B	01-01-2010	8/31/2010
5	I-200-10015-829308	CERTIFIED	04-01-2010	H-1B	09-01-2010	8/31/2013
6	I-200-10244-119864	CERTIFIED	9/17/2010	H-1B	9/17/2010	9/14/2011
7	I-200-09280-855671	CERTIFIED	10/14/2009	H-1B	11-10-2009	2/15/2011
8	I-200-09288-535539	CERTIFIED	10/21/2009	H-1B	11-05-2009	11-04-2010
9	I-200-10189-243755	CERTIFIED	07-09-2010	H-1B	09-01-2010	8/31/2013
10	I-200-10109-451414	CERTIFIED	5/21/2010	H-1B	07-01-2010	6/30/2012
11	I-200-10201-650521	CERTIFIED	7/30/2010	H-1B	09-01-2010	8/31/2013
12	I-200-10201-589228	CERTIFIED	7/26/2010	H-1B	09-01-2010	8/31/2013
13	I-200-10187-139128	CERTIFIED	07-08-2010	H-1B	09-01-2010	8/31/2012
14	I-200-10193-678695	CERTIFIED	7/13/2010	H-1B	09-01-2010	8/31/2011
15	I-200-10209-235134	CERTIFIED	09-07-2010	H-1B	11-05-2010	11-04-2011
16	I-200-10085-608049	CERTIFIED	04-01-2010	H-1B	09-01-2010	8/31/2011

Results per page: 50 1 – 50 of 324687

PERSONAL HISTORY PROJECT HISTORY

REFRESH

# Wrangler Transformation

Cloud Data Fusion | Wrangler

OPERATIONS HUB SYSTEM ADMIN Basic Edition

Cloud Storage Default - mock\_pro\_buck\_1/employment\_onsite\_opportunity.csv

employment\_onsite\_opportunity.csv Columns: 36 | Rows: 1000

Data Insights

Create a Pipeline More

	String case_number	String phase	String render_date	String sponsorship_type	String appointment_start_date	String appointment_end_date	String company_name
1	I-200-09288-936547	CERTIFIED	10/15/2009	H-1B	10/15/2009	10/15/2012	CONCH
2	I-200-09288-275834	DENIED	10/15/2009	H-1B	12-07-2009	12-07-2012	GO CON
3	I-200-09288-149965	DENIED	10/15/2009	H-1B	10/19/2009	10/18/2012	CAPE A
4	I-200-09288-622851	CERTIFIED	10/15/2009	H-1B	10/16/2009	10/15/2012	IBM COI
5	I-200-09288-427508	DENIED	10/15/2009	H-1B	10/30/2009	10/30/2012	SAFE M
6	I-200-09288-381123	CERTIFIED	10/15/2009	H-1B	01-01-2010	12/31/2012	BROWN
7	I-200-09288-220029	CERTIFIED	10/15/2009	H-1B	01-05-2010	01-04-2011	THE PE
8	I-200-09288-924813	CERTIFIED	10/15/2009	H-1B	10/26/2009	10/25/2012	SWISS I
9	I-200-09288-208951	CERTIFIED	10/20/2009	H-1B	10/21/2009	10/21/2012	MCCAN

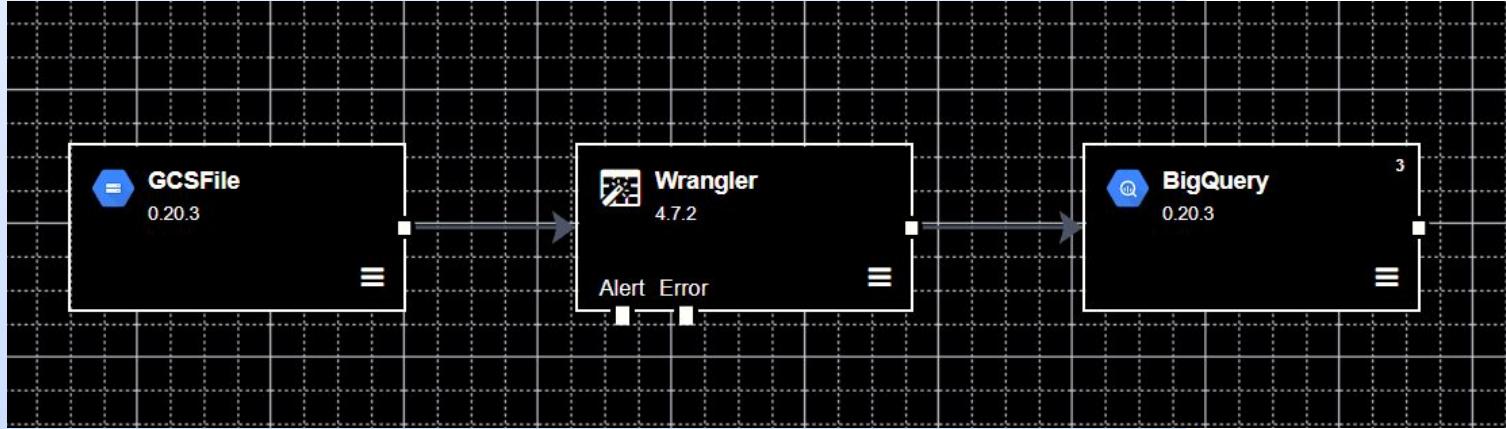
Columns (36) Transformation steps (0)

Search Column names

#	Name	Completion
1	case_number	100%
2	phase	100%
3	render_date	100%
4	sponsorship_type	100%
5	appointment_start_date	100%
6	appointment_end_date	100%
7	company_name	100%
8	company_address	100%
9	company_city	100%
10	company_state	100%
11	company_pincode	100%
12	Classification_code	100%
13	Classification_name	100%

Namespace: default

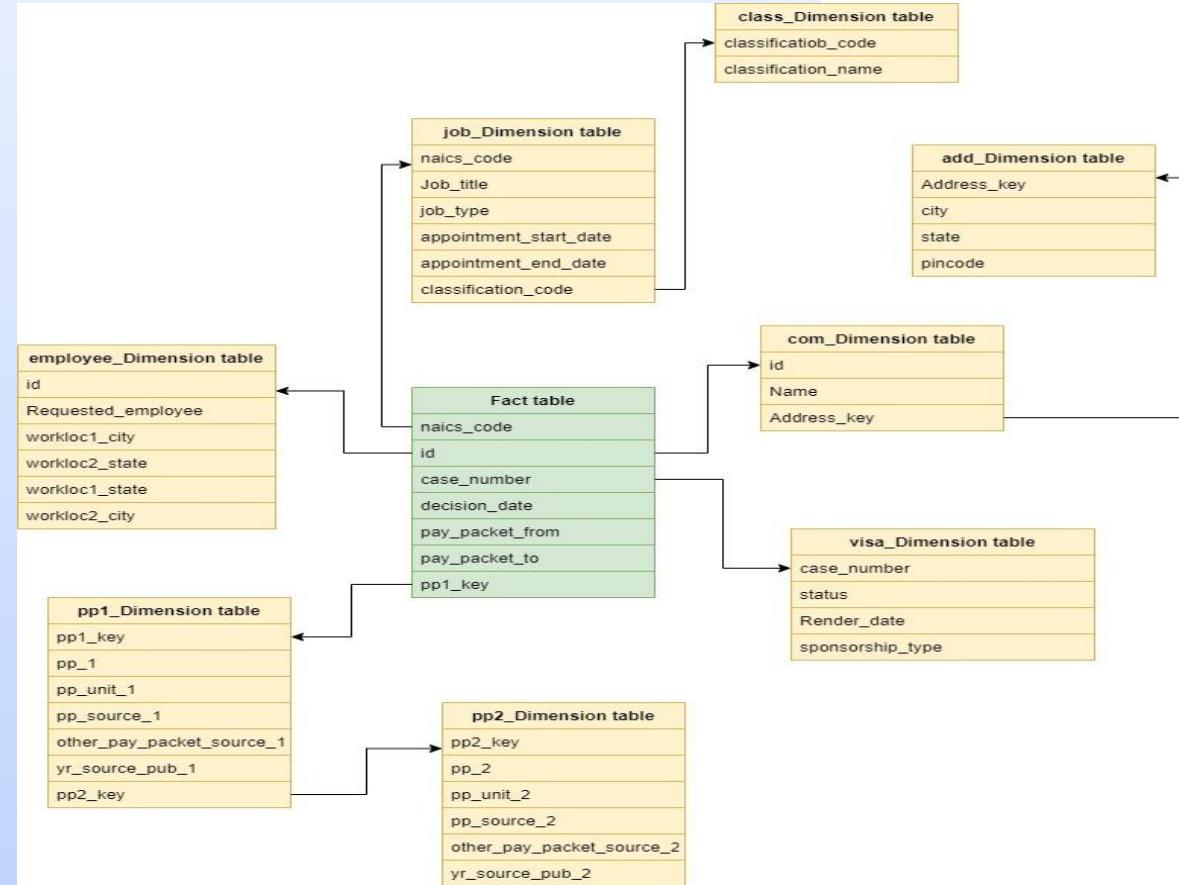
Instance Id: sharp-science-370111/mock-pro-data-fusion



## 4.3

# DATA MODELING

- Fact table contains the measuring of the attributes of a dimension table.
- Dimension table contains the attributes on that truth table calculates the metric.





tableau

- Tool used: Tableau
- Tableau is a Business Intelligence tool for visually analyzing the data.
- Connect Tableau with Google BigQuery
- Perform visualization
- Data Visualization with tableau is nothing but the process of presenting information through visual rendering.



## Connections

Add

BigQuery  
Google BigQuery

## Billing Project

My First Project

## Project

My First Project

## Dataset

mock\_pro\_dataset

## Table

mock\_pro\_tb

New Custom SQL

New Union

New Table Extension

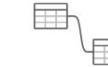
 Use Legacy SQL

Go to Worksheet

## mock\_pro\_tb (mock\_pro\_dataset)

Connection  
 Live  Extract

mock\_pro\_tb



Need more data?

Drag tables here to relate them. [Learn more](#)

mock\_pro\_tb 36 fields 324687 rows

100

## Name

mock\_pro\_tb

## Fields

Type	Field Name	Physical Ta...	Rem...
Abc	Case Number	mock_pro_tb	case_...
Abc	Phase	mock_pro_tb	phase
Abc	Render Date	mock_pro_tb	rende...

Abc	Abc	Abc	Abc	Abc
Case Number	mock_pro_tb	mock_pro_tb	mock_pro_tb	mock_pro_tb
Phase	Phase	Phase	Render Date	Sponsorship Type
I-200-09334-520385	CERTIFIED	11/30/2009	H-1B	12-11-2009
I-200-10223-221072	WITHDRAWN	08-11-2010	H-1B	09-01-2010
I-200-10228-922855	CERTIFIED	8/16/2010	H-1B	09-01-2010
I-200-09315-469049	CERTIFIED	11/20/2009	H-1B	01-01-2010
I-200-10015-829308	CERTIFIED	04-01-2010	H-1B	09-01-2010
I-200-10244-119864	CERTIFIED	9/17/2010	H-1B	9/17/2010

 Data Source

Sheet1



# DATA DICTIONARY

## LCA

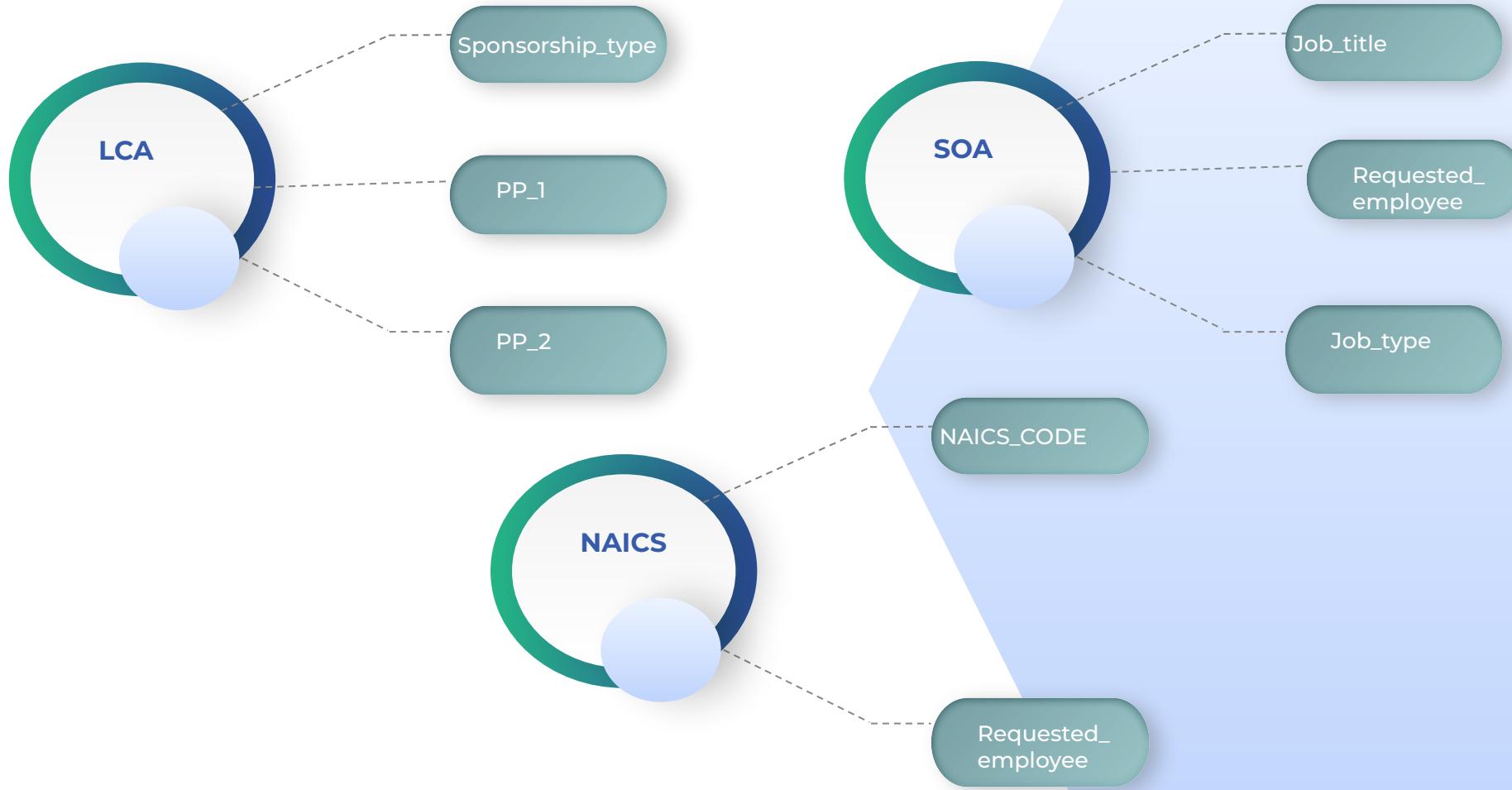
The Labor Condition Application (LCA) is an application filed by prospective employers on behalf of workers applying for work authorization for the non-immigrant statuses H-1B, H-1B1 (a variant of H-1B for people from Singapore and Chile) and E-3 (a variant of H-1B for workers from Australia).

## SOC

The Standard Occupational Classification (SOC) System is a United States government system of classifying occupations. It is used by U.S. federal government agencies collecting occupational data, enabling comparison of occupations across data sets.

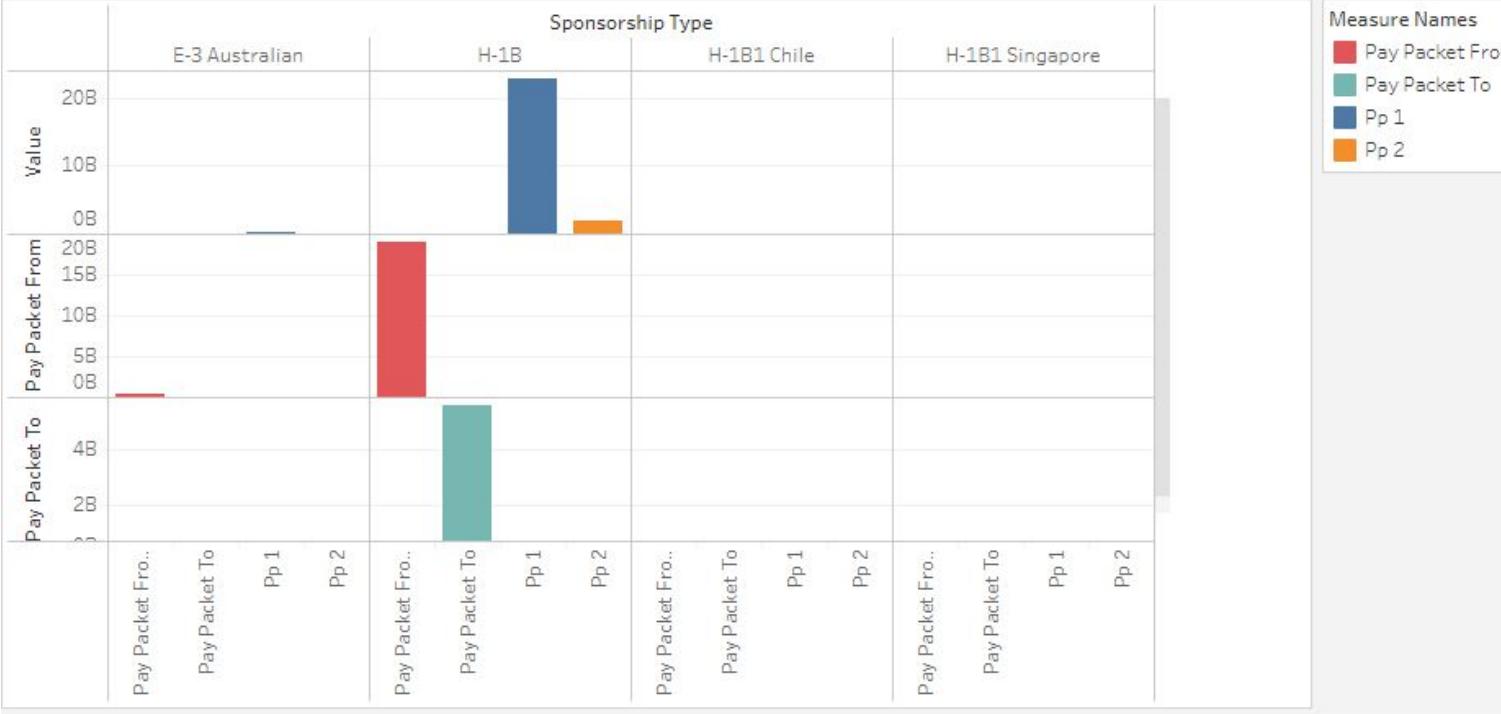
## NAICS

The North American Industry Classification System (NAICS) is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy.



iii Columns	Sponsorship Type	Measure Names
-------------	------------------	---------------

iii Rows	Measure Values	SUM(Pay Packet From)	SUM(Pay Packet To)
----------	----------------	----------------------	--------------------



LCA

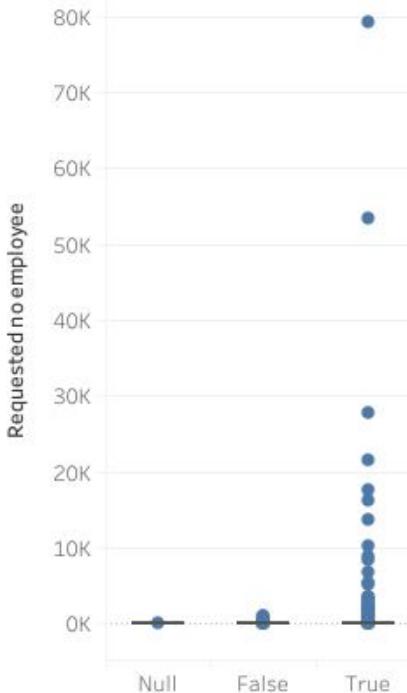
iii Columns

Job Type

Rows

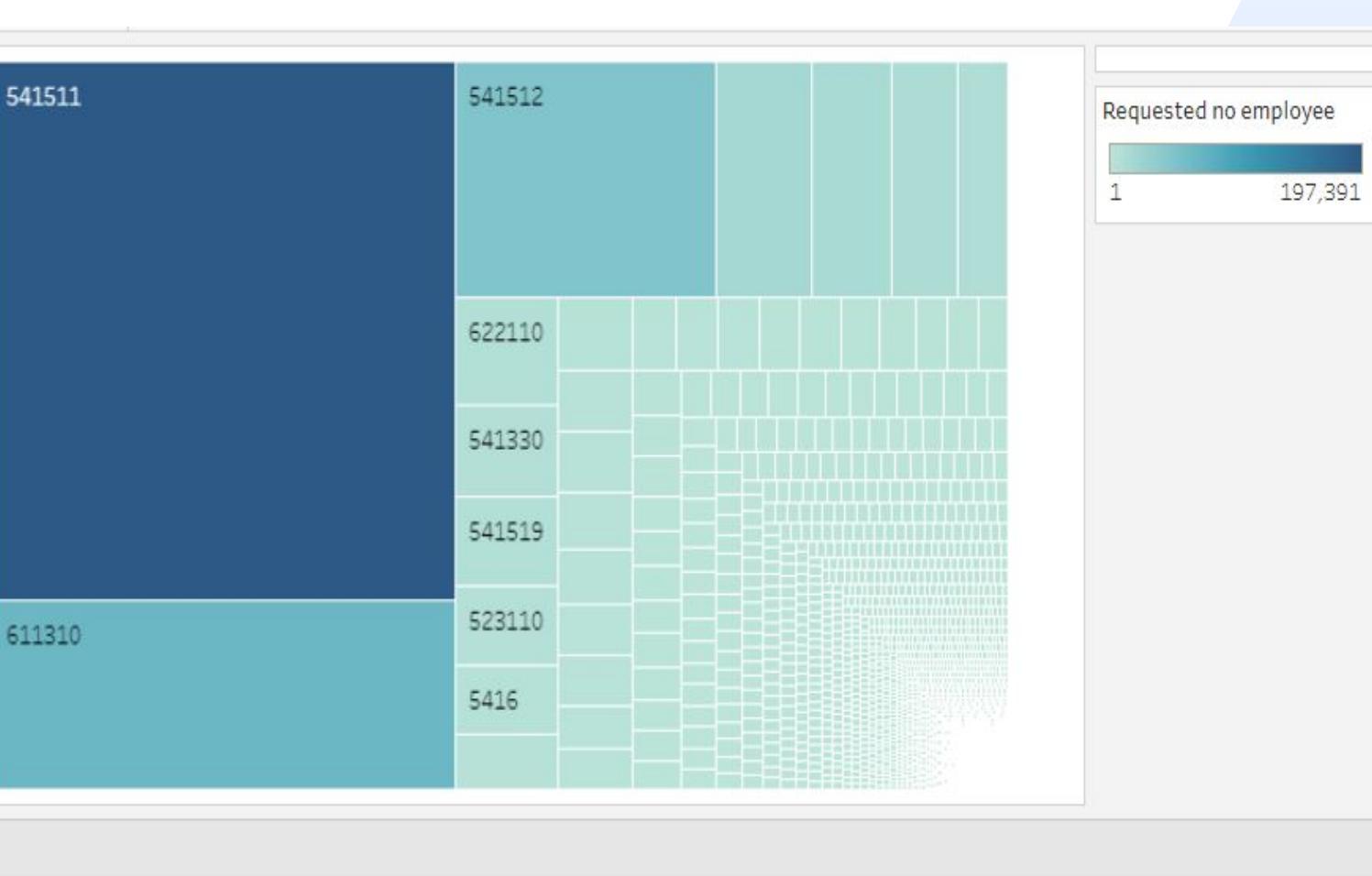
SUM(Requested no e...)

Job Type



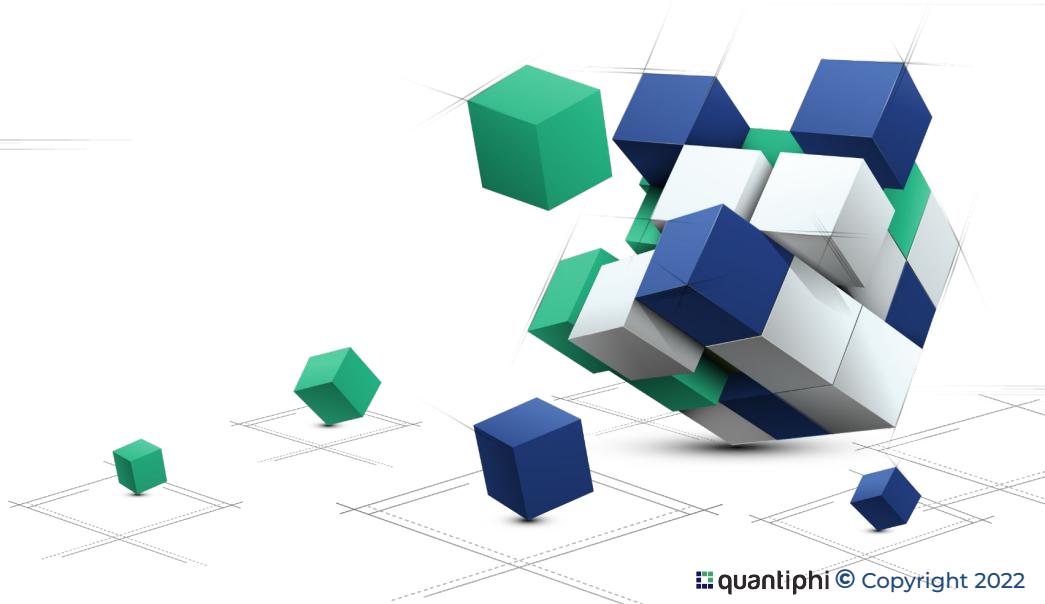
SOA

# North American Industry Classification System



# 04

## CONCLUSION



Our approach to this project is as follows:

- Gather data understanding via data dictionary and domain knowledge.
- Then clean the raw data using pandas MS Excel, GCP data prep.
- Perform ETL using GCP.,
- Connect Tableau with Google BigQuery
- Perform visualization
- Data Visualization with tableau is nothing but the process of presenting information through visual rendering.

# THANK YOU!

