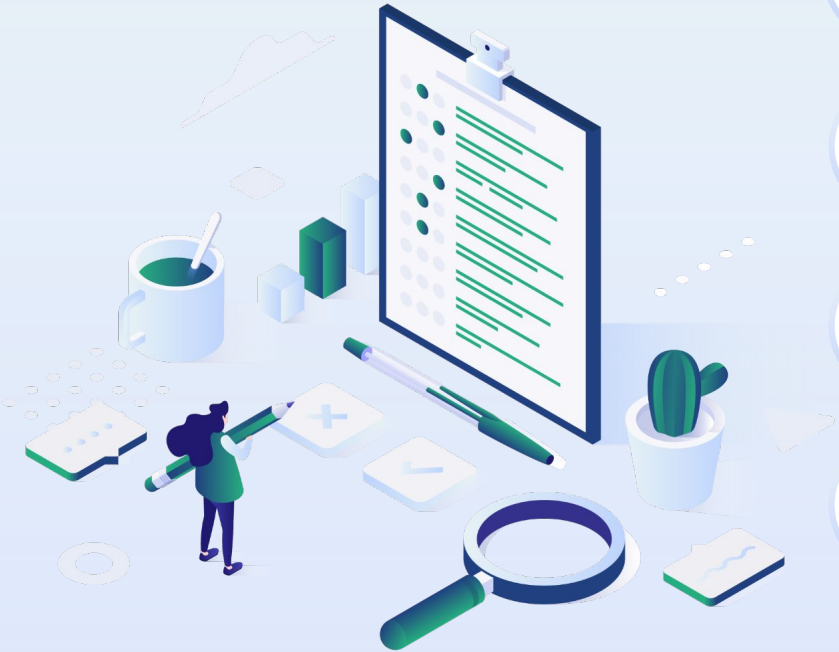
EMPLOYEE ONSITE OPPURTUNITY

quantiphi
AI - Data - Cloud

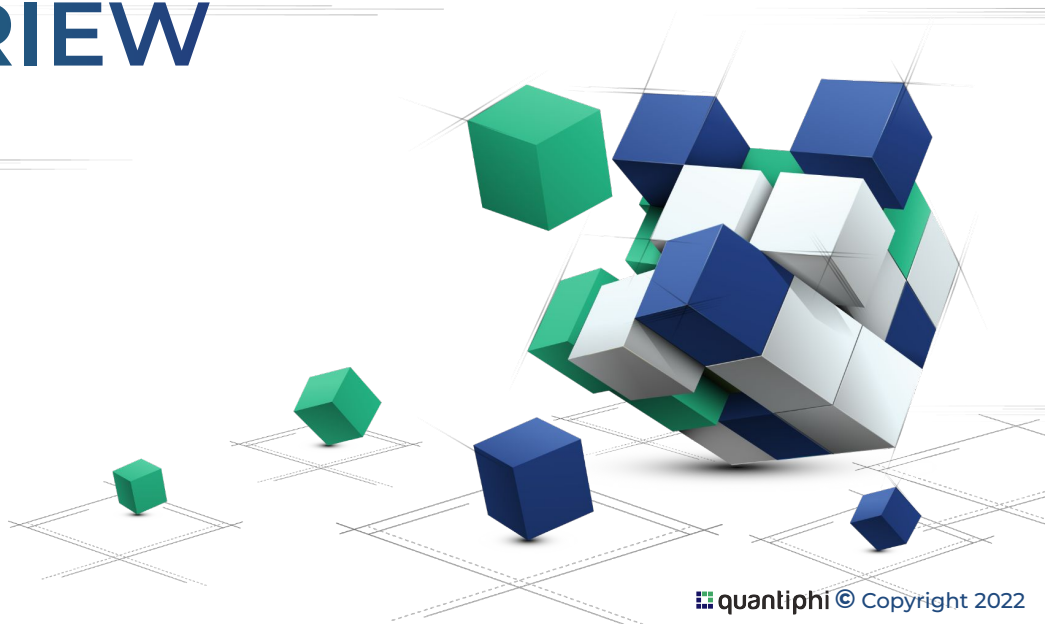
01 Agenda



- PROJECT OVERVIEW
- WORKFLOW
- IMPLEMENTATION
- CONCLUSION

01

PROJECT OVERIEW



01

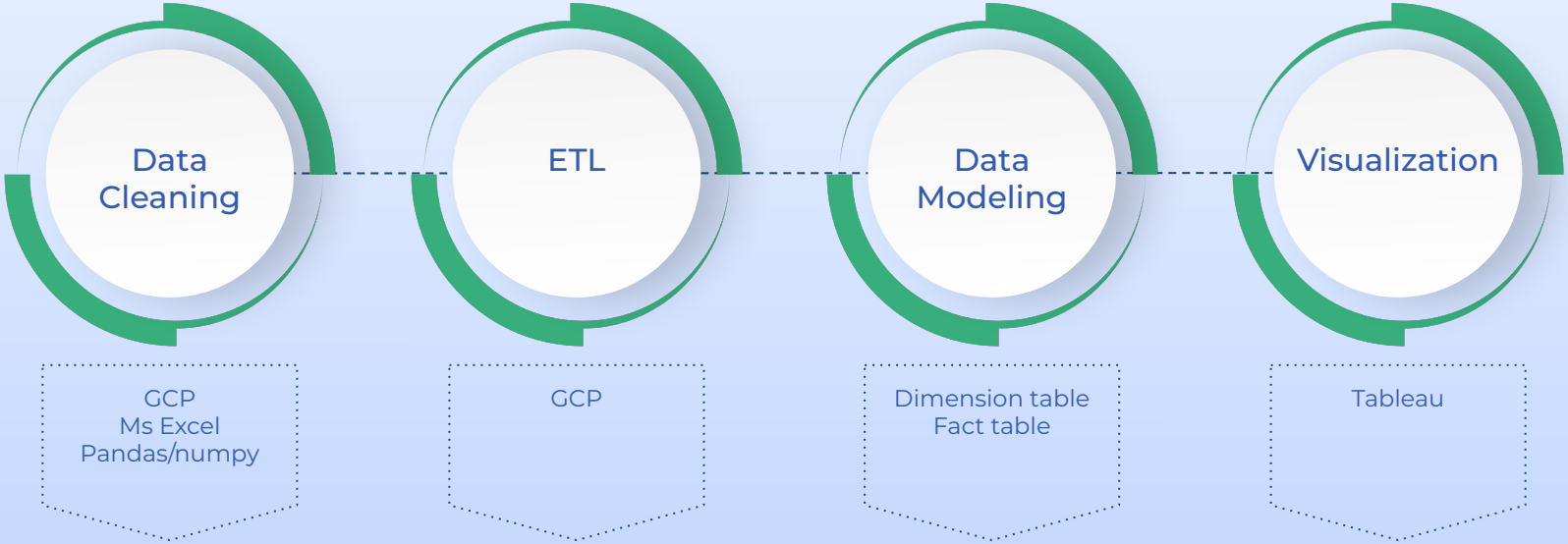
PROJECT OVERVIEW

- 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.

quantiphi
AI - Data - Cloud

02
WORK FLOW

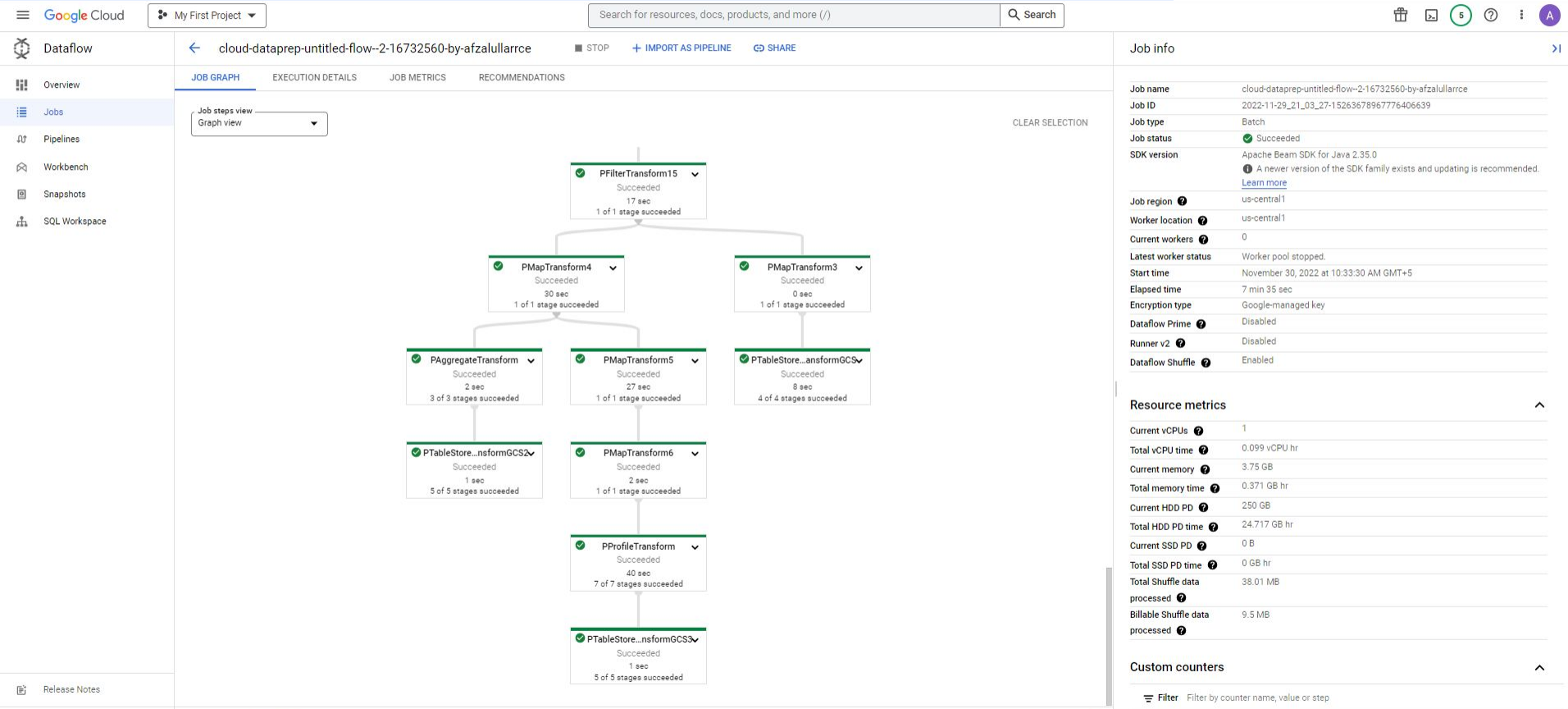
WORK FLOW



03
IMPLEMENTATION

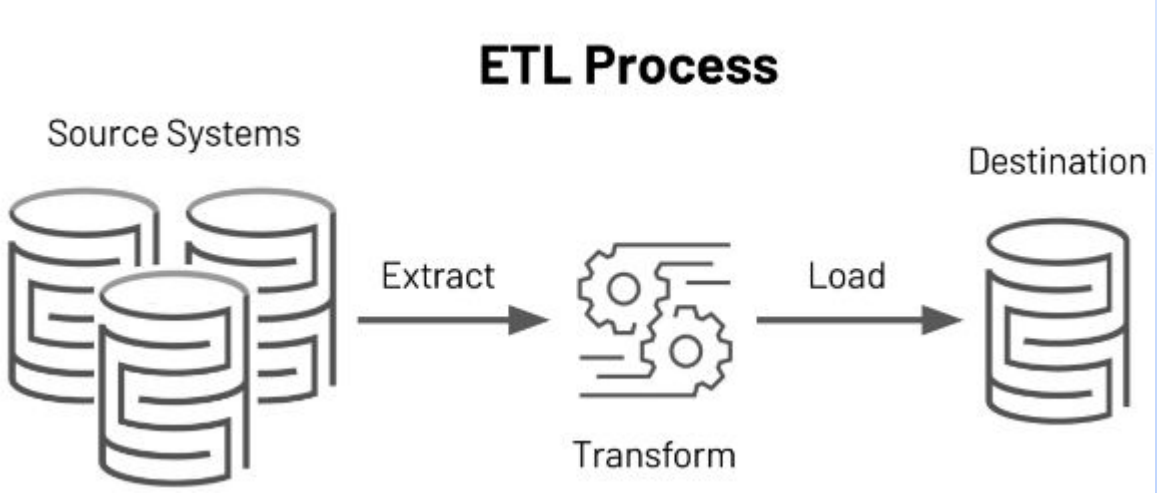
3.1 DATA CLEANING

- 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



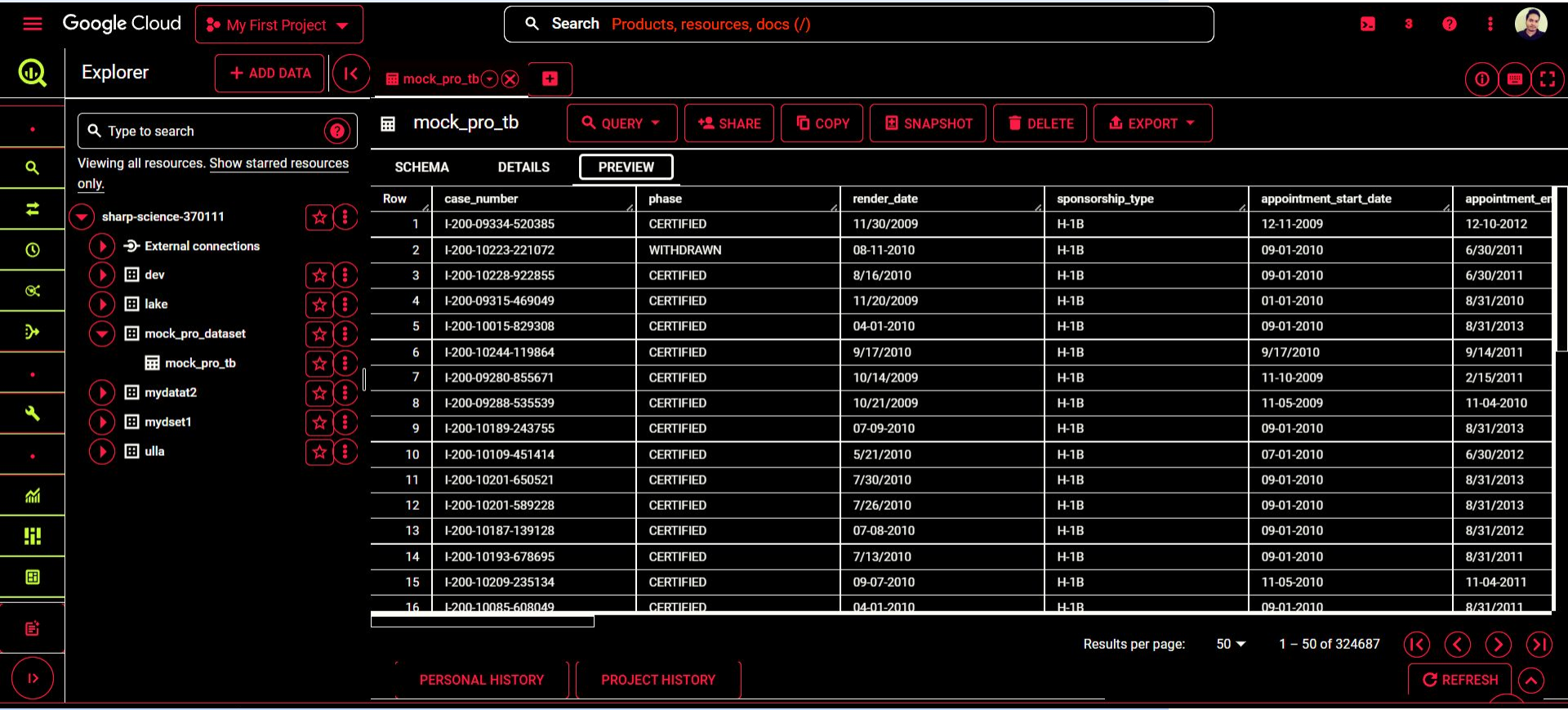
4.2 EXTRACT-TRANSFORM-LOAD

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.

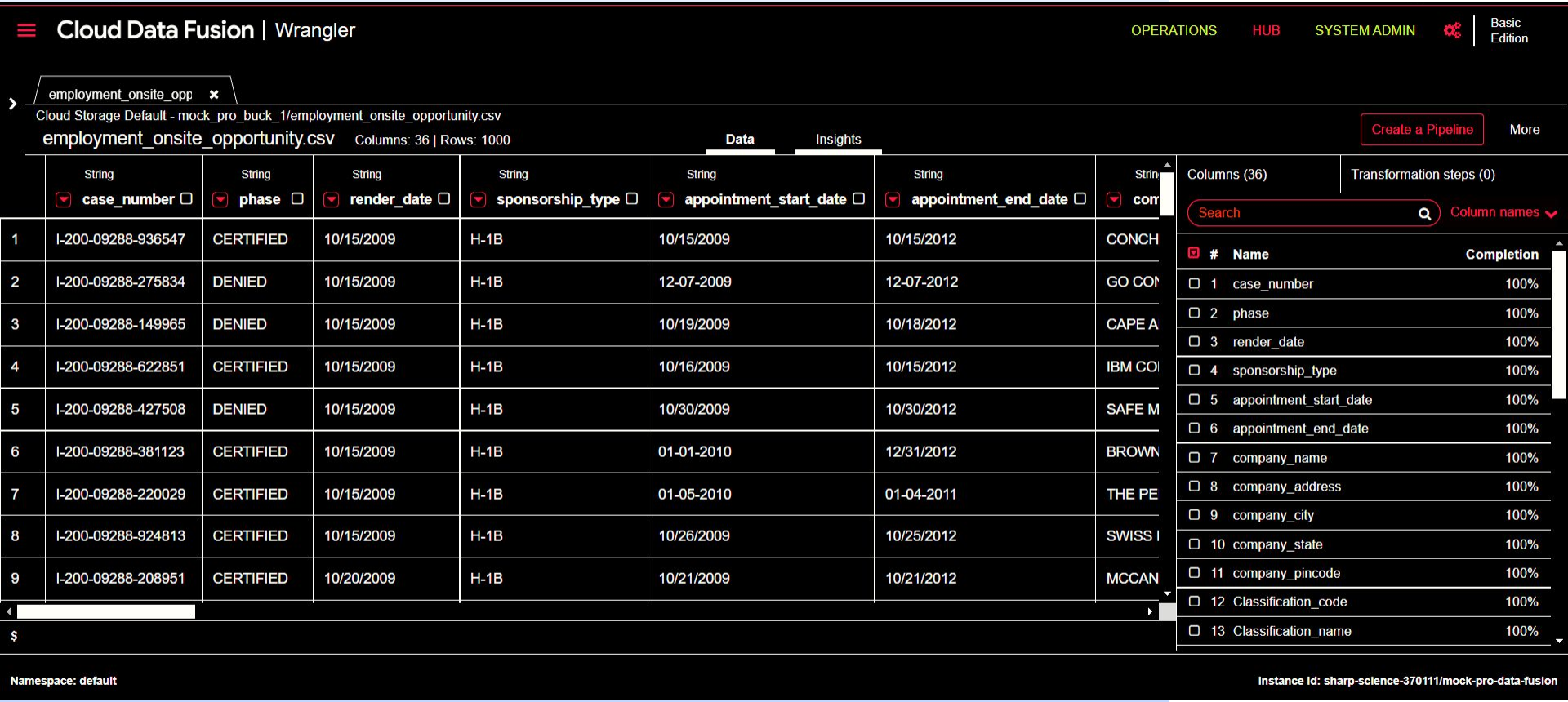


1) Create a bucket by CLI
    a) gcloud storage buckets create gs://mock_pro_bucket1
2) Import cleaned CSV file to the bucket
3) Extract this into Wrangler
4) Wrangler - Transform
    a) Build a robust batch data pipeline using data fusion with wrangler
5) Load this transformed data into BigQuery(Data warehouse)

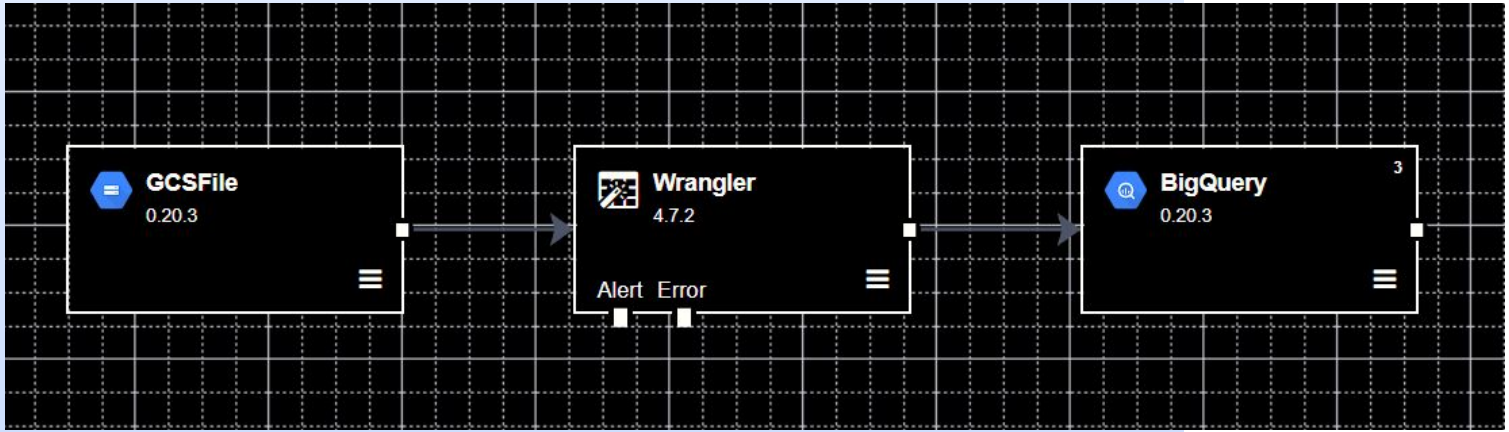
Data warehouse- BigQuery



Wrangler Transformation

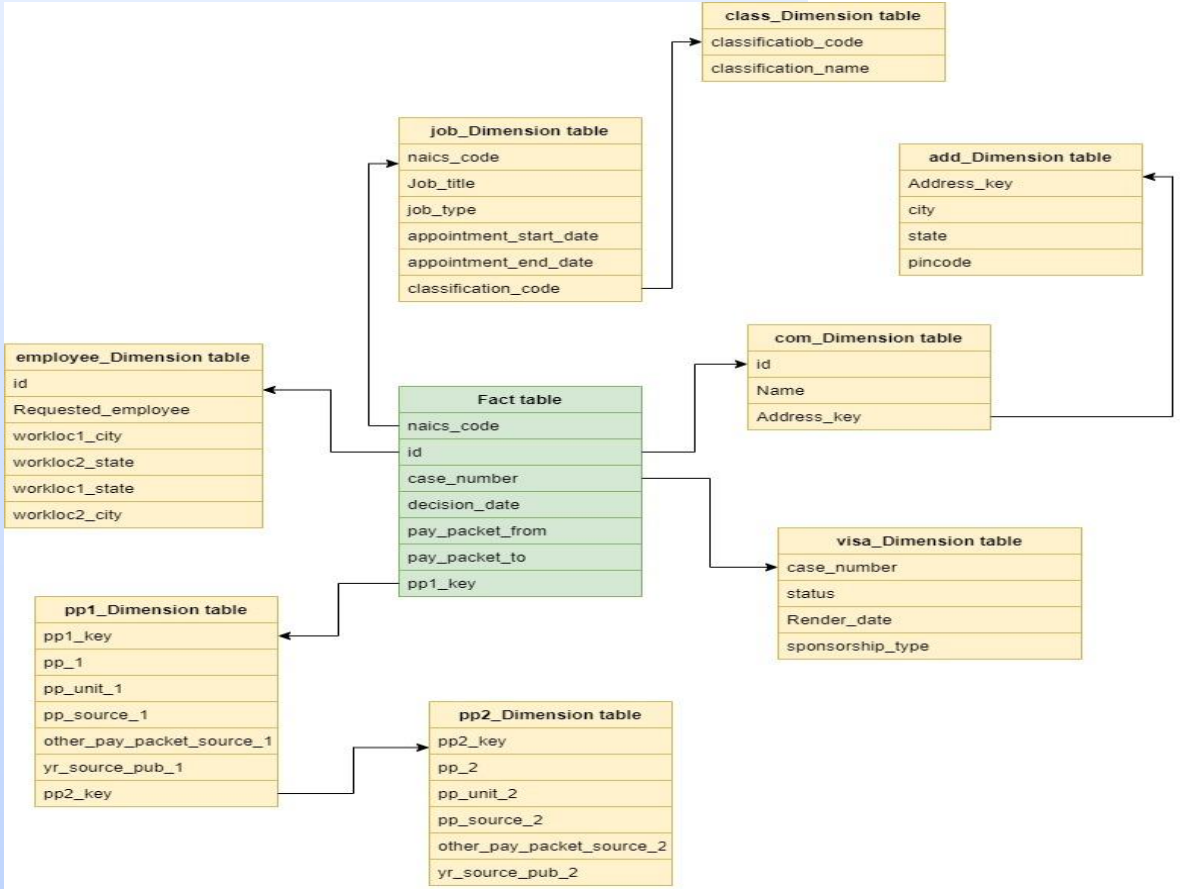


	String case_number	String phase	String render_date	String sponsorship_type	String appointment_start_date	String appointment_end_date	Strin com
1	I-200-09288-936547	CERTIFIED	10/15/2009	H-1B	10/15/2009	10/15/2012	CONCH
2	I-200-09288-275834	DENIED	10/15/2009	H-1B	12-07-2009	12-07-2012	GO CON
3	I-200-09288-149965	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



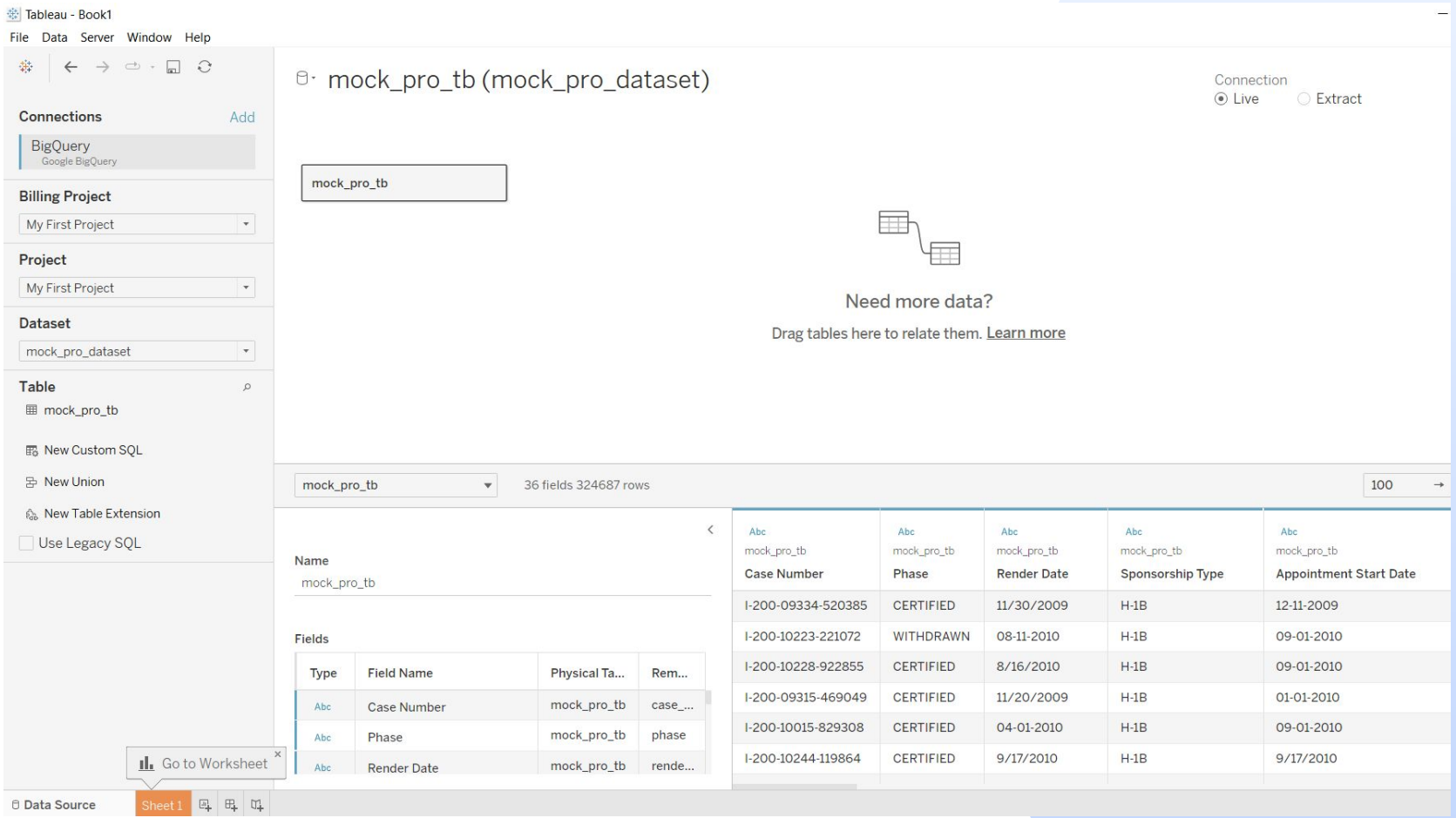
4.3 DATA MODELING

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



4.4 VISUALIZATION

- 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.

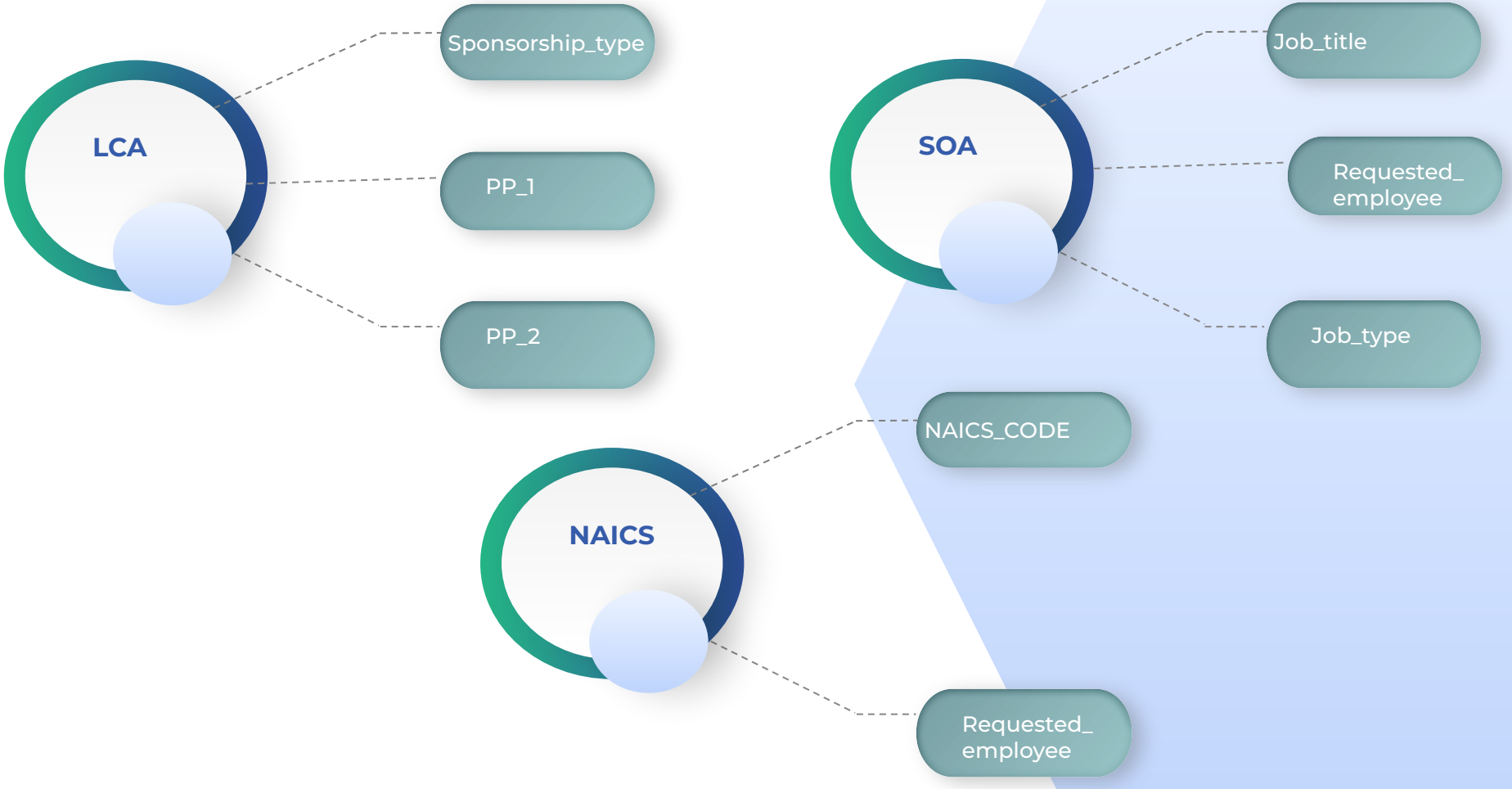


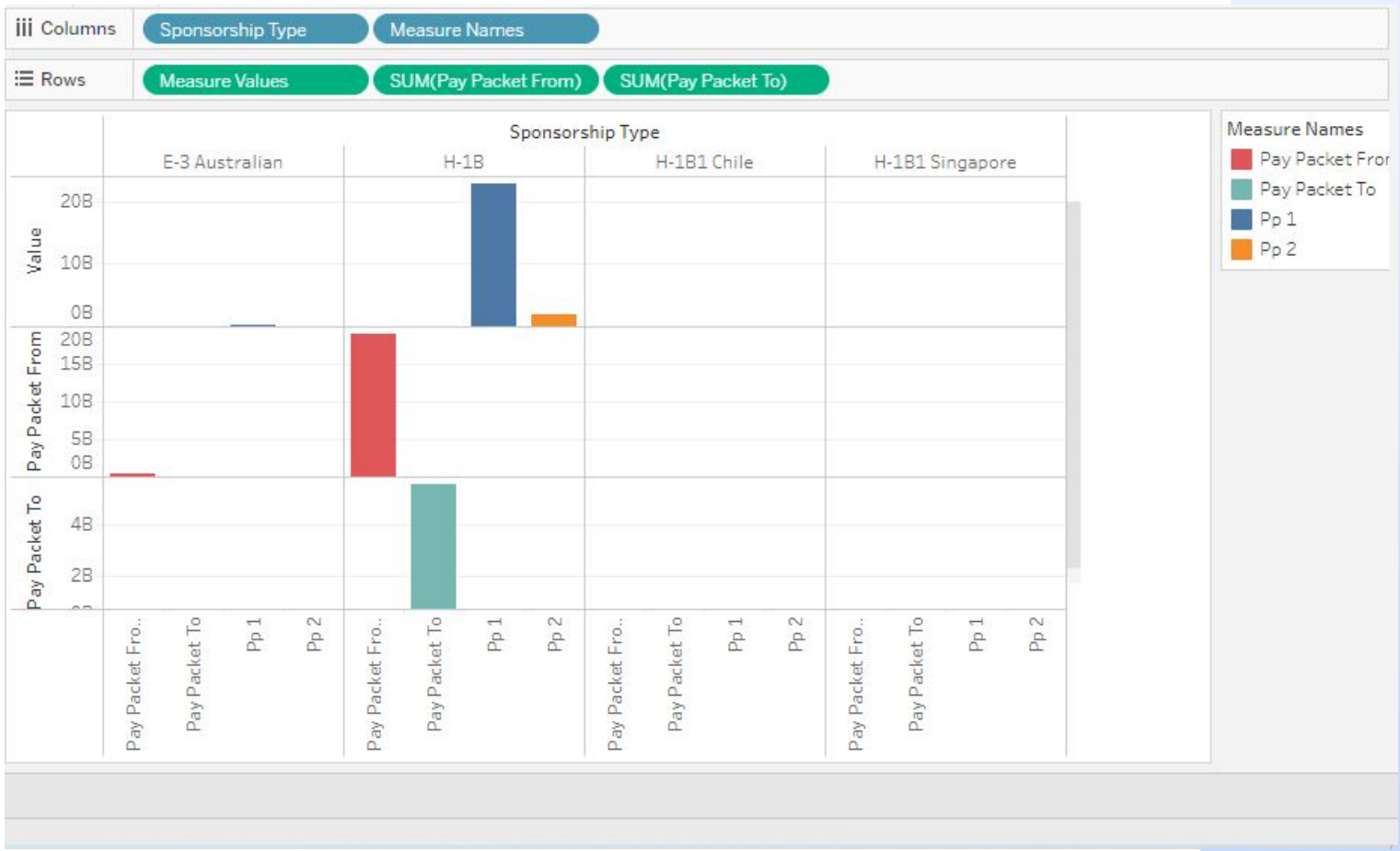
DATA DICTIONARY

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).

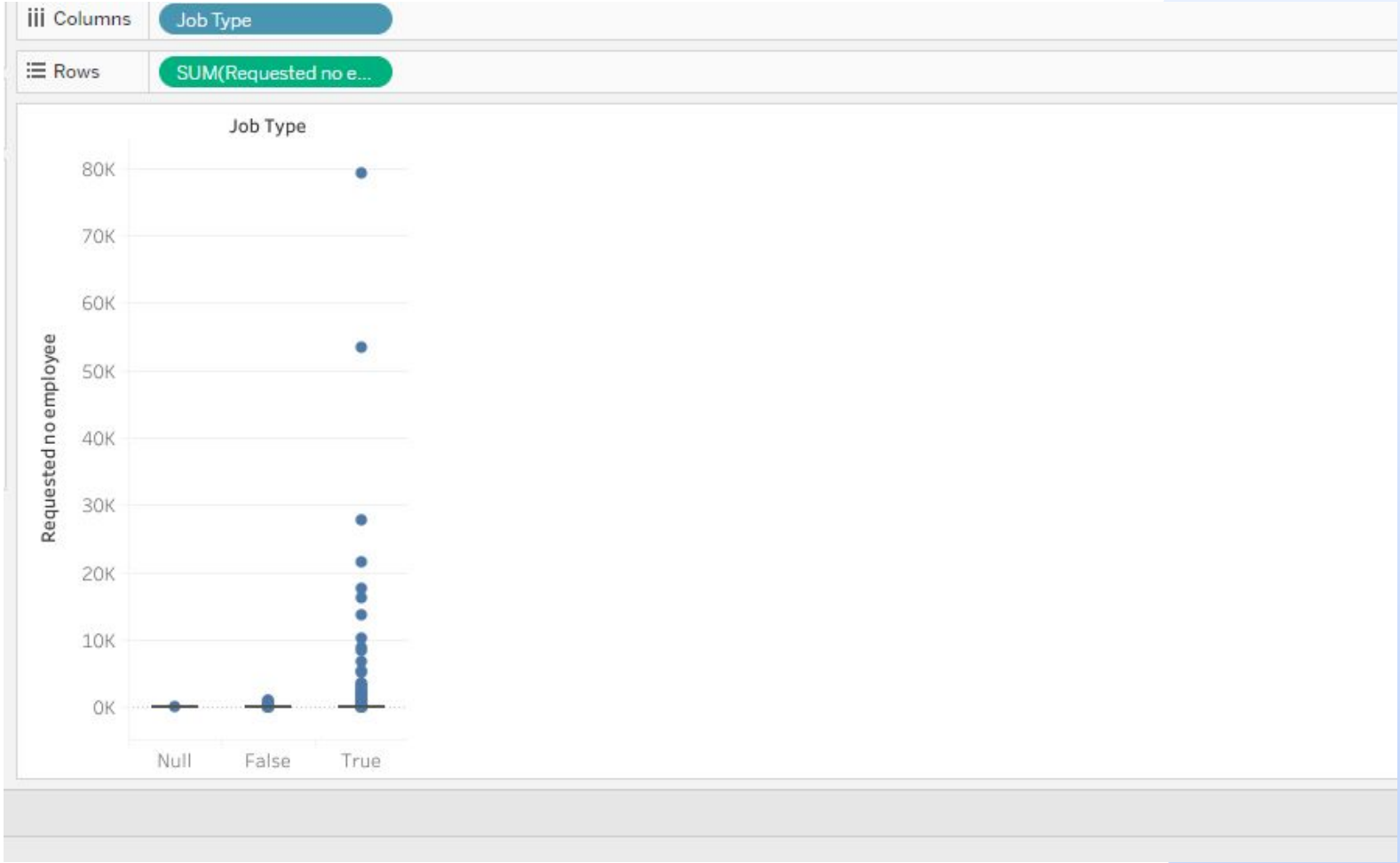
SOC
The Standard Occupational Classification (SOC) System is a United States government system of classifying occupations. It is used by U.S. federal government agencies collecting occupational data, enabling comparison of occupations across data sets.

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.

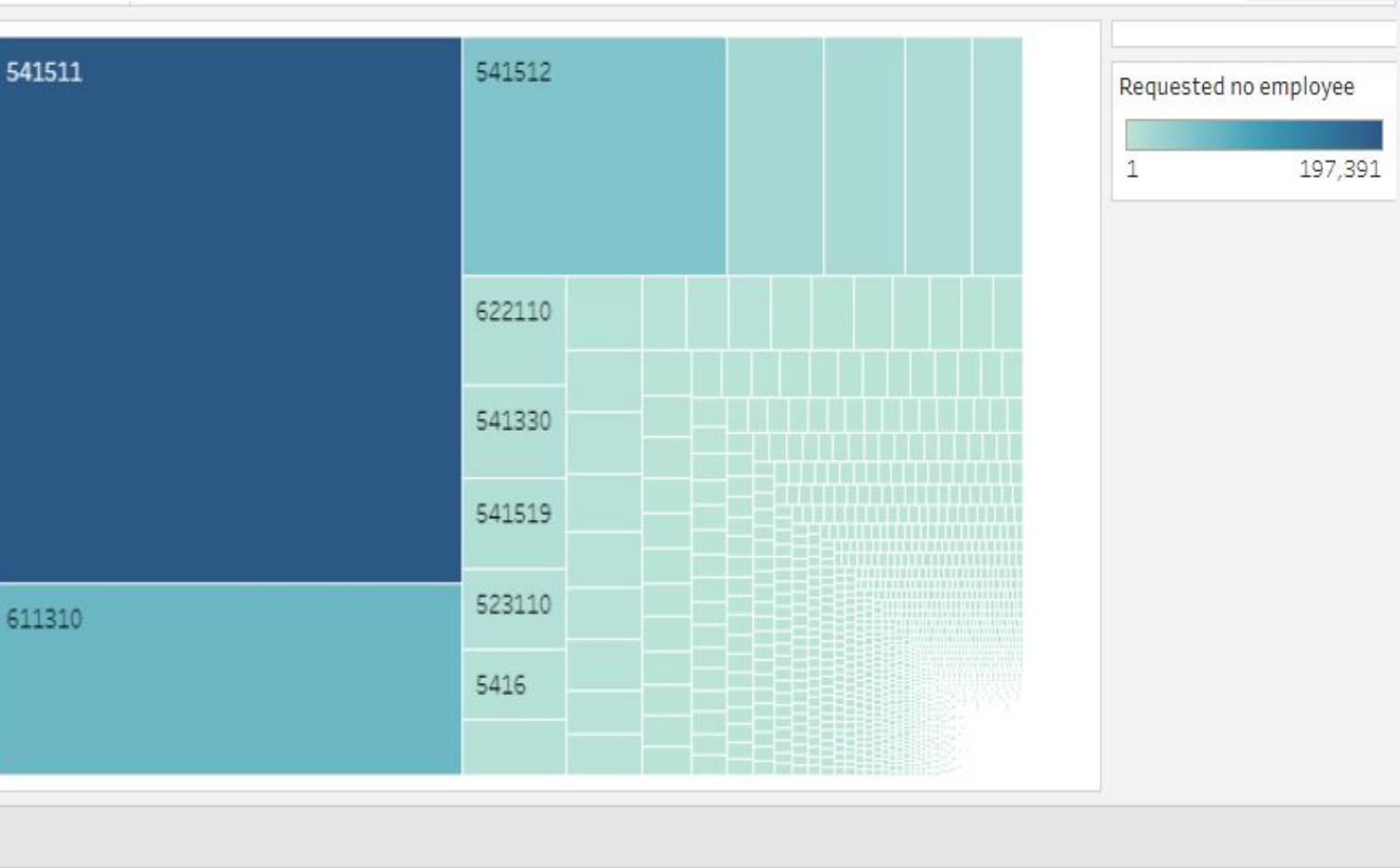




LCA

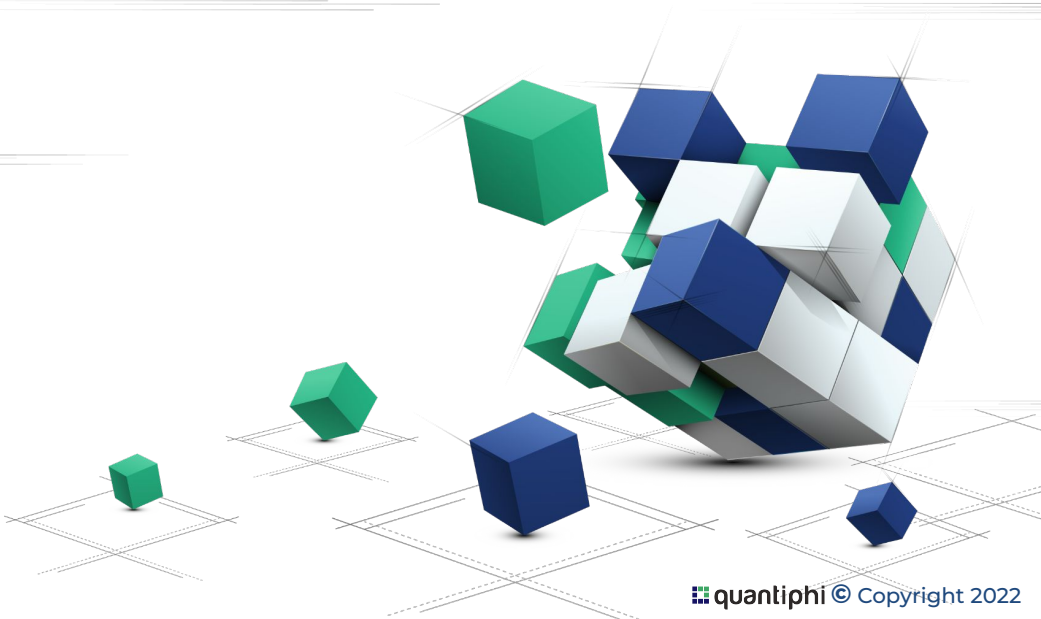


SOA



North American Ind
Classification Sys

04
CONCLUSION



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!