# Climate Data Stack

Reflections from SALURBAL and planning for P20/SALURBAL-climate.

Before we start lets come to an agreement on the terms we will use to describe data processes we will be discussing. Here are some key vocabularies:

* Vocabulary sources:
  + [Dagster Data Engineering Terms explained glossary](https://dagster.io/glossary)  have very accessible definition for key data engineering terms that non-engineers can easily understand.
* Data Engineering Terms
  + [**ETL**](https://dagster.io/glossary/etl) is just moving raw data into another data system. This is the hardest and probably most important part of what we do. Some examples of ETL pipelines include:
    - Raw -> Working group level (Mortality, BEC, SEC)
    - Working group level -> Project level (SALURBAL Portal)
  + [**Denormalization**](https://dagster.io/glossary/denormalize)**:** is the process of optimizing data structures into something tidy to make data more ergonomic by reducing number of joins needed to do analytics.

# Introduction

There are a few key components to a data stack:

* Raw data
* ETL
* Data warehouse
  + Database engine
  + Orchestration framework
* BI Tool

I Think in SALURBAL 1 what we had were

* Individual working groups doing ETL from raw\_data to cleaned\_data
* Then renovations which a centralized group ETL cleaned\_data into a centralized data warehouse
* Then a public facing

There are three major points of improvement for this pipeline.

* Data provenance was not something we built.
  + The ETL pipeline was just built to inventory existing data and not built to handle transformations.
* Dataset level metadata
* Use friendly data entry