

# The Proactive Data Catalog

June 2022

**stemma**<sup>TM</sup>

from the creators  
of amundsen

# Introduction



## Bálint Haller

Senior Software Engineer, Stemma

Previously Data Engineering @  
Shapr3D

[www.linkedin.com/in/balinhaller](https://www.linkedin.com/in/balinhaller)

# Proactive vs reactive

## Proactive

acting in anticipation  
of future problems,  
needs, or changes

## Reactive

readily responsive to  
a stimulus

# Proactive

Anticipate changes

- data freshness
- data structure
- meaning

# Reactive



# Proactive

Anticipate changes

- data freshness
- data structure
- meaning

# Reactive

Observe changes, and act if the need arises

- Fix broken ETL/ELT pipelines
  - calculations
  - meaning

**How does this connect  
to metadata?**

**stemma™**

# Data Catalog as a thing

- An organised inventory of data assets in the organisation
- Helps data professionals collect, organise, access, and enrich metadata to support data discovery and governance





# Data Catalog as a thing

- Central part is the **description** of datasets, columns, dashboards
- And any other data asset potentially
- Slightly different from ‘Data Observability’ tools that are meant to detect anomalies
- Can be integrated though!

Q|

×

[Browse](#)
[Glossary](#)
[Help & Docs](#)

BH

[Admin](#)

open\_data.case\_demographics\_age

[Tables](#) • [Snowflake](#) • [ca\\_covid](#)

☆

[Airflow](#)
[Github](#)

Lineage graph

Preview

Message

Status

[Set Status](#)

Description

Age categories of COVID-19 cases and deaths as reported by local health departments. This includes:

- positive cases
- deaths
- testing results

Issues

P1

SD-31

Wrong column name

To Do

P1

SD-9

There's an issue with this table

To Do

P1

SD-32

Data is out of date

Done

P1

SD-7

This table is busted

Done

View all 4 issues

|

[Report an issue](#)

Last Updated

Dec 06, 2021 7pm CET

Owners

B

bob@stemma.ai

DJ

Dorian Johnson

Columns (7)

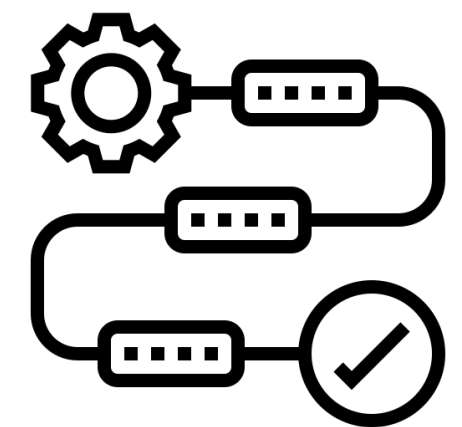
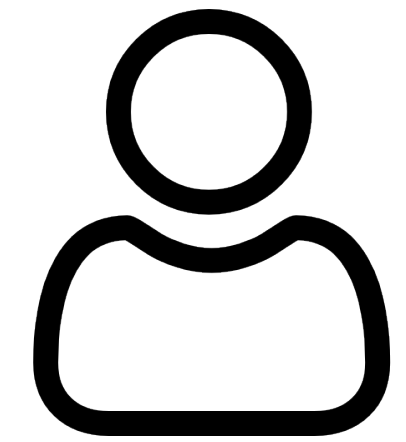
Dashboards (0)

Sort by ▾

NAME	TYPE	BADGES
<div>▾</div> <div>ca_percent</div> <div>Percent of age in relation to overall population.</div>	float	⋮
<div>▾</div> <div>age_group</div> <div>Count of covid cases that result in deaths.</div>	text	<div>npi</div> <div>pii</div> <div>⋮</div>
<div>▾</div> <div>deaths</div> <div>Cumulative number of COVID-related deaths as reported by local health departments.</div>	number	⋮
<div>▾</div> <div>case_percent</div> <div>Percent of total cases.</div>	float	⋮
<div>▾</div> <div>totalpositive</div> <div>Cumulative number of COVID confirmed cases as reported by local health departments.</div>	number	<div>Certified</div> <div>⋮</div>
<div>▾</div> <div>deaths_percent</div> <div>Percent of total deaths.</div>	float	⋮
<div>▾</div> <div>date</div> <div>Date reported.</div>	date	<div>partition</div> <div>⋮</div>

# Data Catalog as a thing

- General shift from people-oriented to automated
  - This is mostly in the **reactive -> proactive** direction
  - Automated **doesn't lose the human element**
  - Dataset **owners maintain** their own descriptions (instead of data stewards)
- Supporting tools in the **modern data stack**
- **connecting things** rather than trying to pull in every feature (SQL editor, dashboards, etc.)



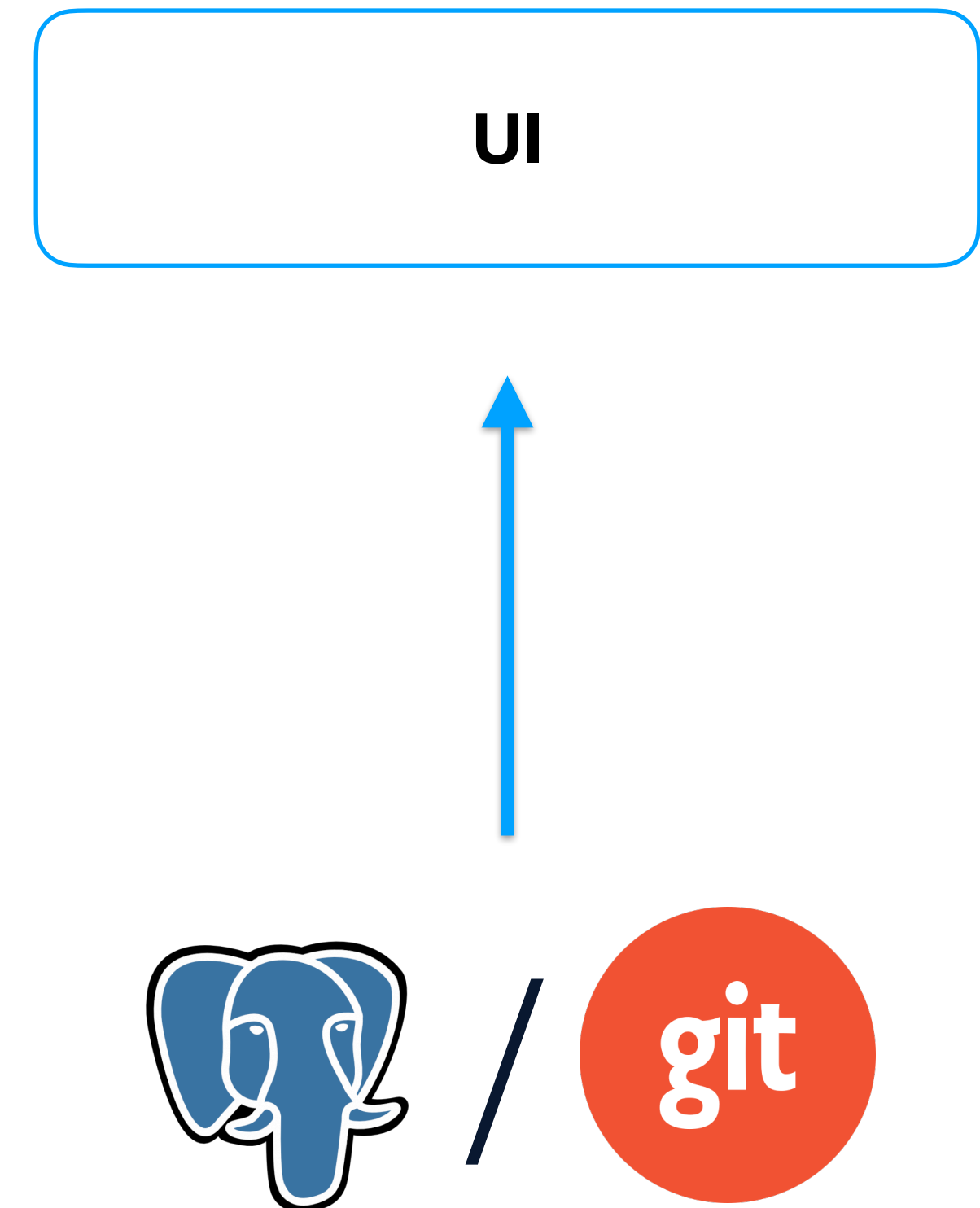
The background is a dark navy blue. It features several abstract, organic shapes in a lighter blue color. In the top right, there is a large, rounded, teardrop-like shape. To its right is a solid blue circle. In the bottom left, there is another solid blue circle and a large, rounded, teardrop-like shape. The text is centered in the middle of the frame.

# How do you handle descriptions?

stemma™

# Handling descriptions

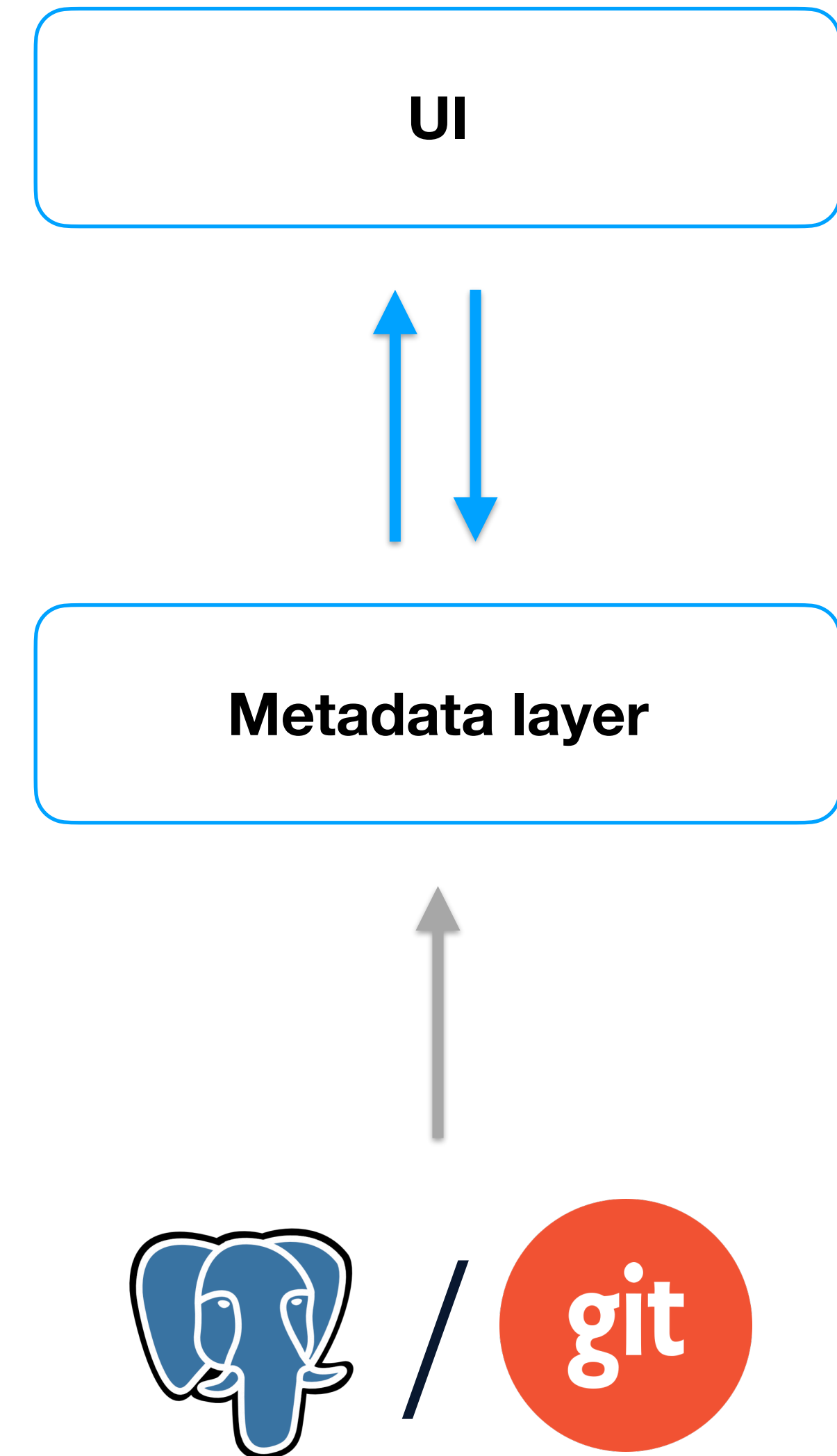
- From the DB/event schema
- The closer to source, the better\*
- Can even get rid of the metadata layer and serve the exact same descriptions





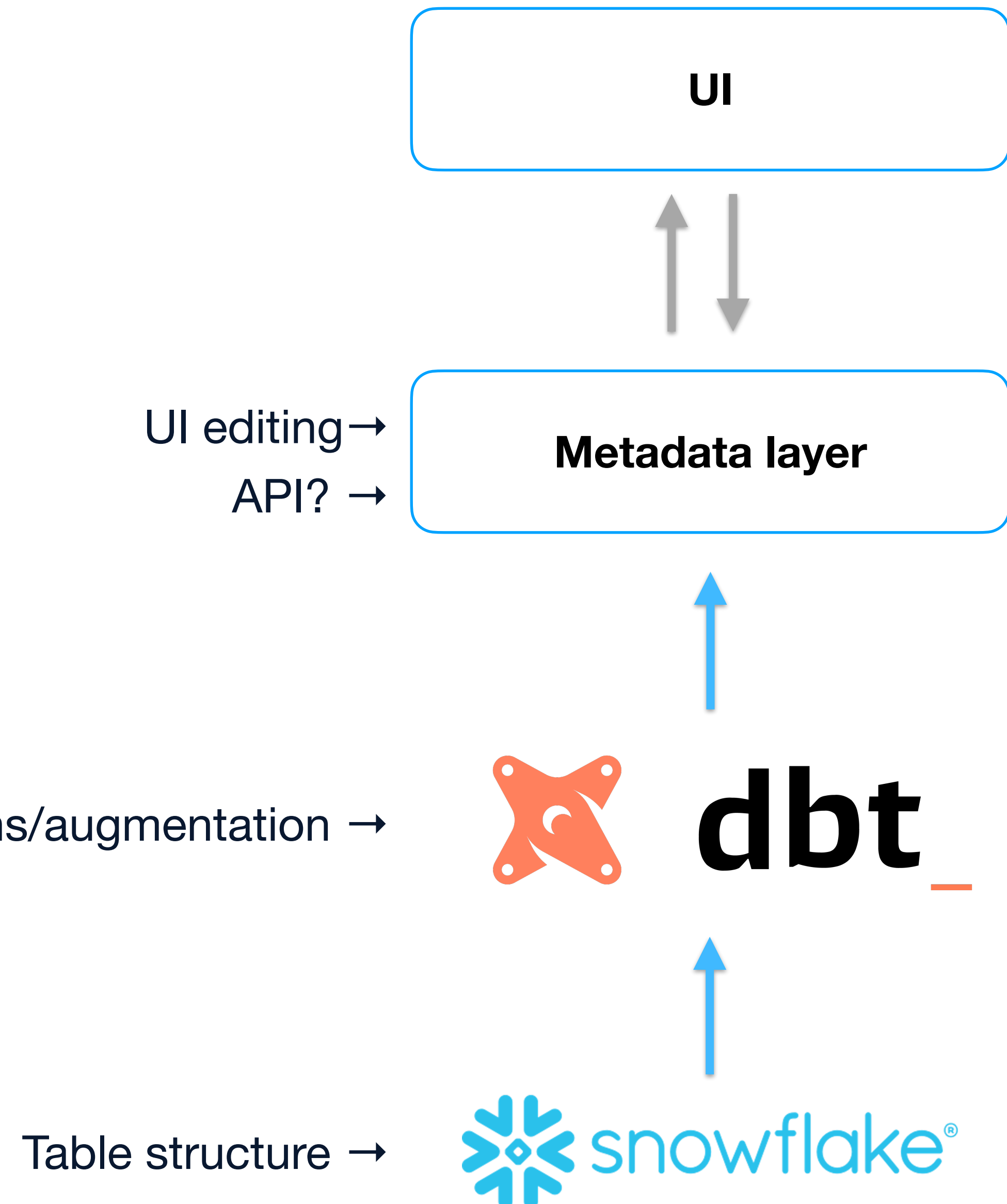
# Handling descriptions

- However! You **still need UI editability** for non-technical users
- Otherwise you run the risk of not having a 'network effect'
- Easier to get out of sync
- Can be very reactive and outdated



# Handling descriptions

- Some companies have extra layers
- E.g. DBT on top of a data warehouse
- Need to be able to resolve these
- Some catalogs expose an API
- Ingestion or augmentation
- Very flexible, but needs infra





**What to do?**

# Handling descriptions

## 1.Display everything

- More complex
- Not the right call if you need canonical descriptions

## 2.Establish a resolution strategy

### 1.Static priority list

### 2.Based on freshness

### 3.Can get fancy here, but it comes at the cost of transparency

### Description

Age categories of COVID-19 cases and deaths as reported by local health departments. This includes:

- positive cases
- deaths
- testing results

### Issues

P1	SD-31	Wrong column name	To Do
P1	SD-9	There's an issue with this table	To Do
P1	SD-32	Data is out of date	Done
P1	SD-7	This table is busted	Done

[View all 4 issues](#) | [Report an issue](#)

### Last Updated

Jun 13, 2022 7pm CET

### Owners

 bob@stemma.ai



# Surfacing missing descriptions

- Do you surface missing descriptions?
- Useful when the proactive approach fails, and you need to augment a significant portion of your descriptions by hand
- or just to clean up a couple outlier datasets

## Description

Age categories of COVID-19 cases and deaths as reported by local health departments. This includes:

- positive cases
- deaths
- testing results

## Issues

P1	SD-31	Wrong column name	To Do
P1	SD-9	There's an issue with this table	To Do
P1	SD-32	Data is out of date	Done
P1	SD-7	This table is busted	Done

[View all 4 issues](#) | [Report an issue](#)

## Last Updated

Jun 13, 2022 7pm CET

## Owners

 bob@stemma.ai

**How do you handle  
changes?**

**stemma™**

# How do you handle changes?

- Changes can be painful, especially if multiple layers have information about your datasets' structure
- Lineage becomes important
  - Helps with notifying downstream and assessing impact
  - Ability to message downstream dataset owners can help prepare (more automation isn't usually possible)



# Handling changes

- Shared term glossary is helpful – Proactive
  - Leaves less room for errors
- Column and table description linking
- Anything to make your descriptions DRY
- Data freshness and size alerts
  - The realm of observability
- Deleting stale data

## Cases

Ungrouped

### Group

Ungrouped

### Last Updated

Apr 19, 2022 6pm CEST

### Owners

 Grant Seward  Mark Grover  Merilin Pisina

### Related Columns

 [open\\_data.case\\_demographics\\_ethnicity.cases](#)

 [open\\_data.statewide\\_cases.totalcountconfirmed](#)

### Definition

COVID cases as defined by CDC for state of California.

**Related tables:** [open\\_data.statewide\\_cases](#)



# Finding and fixing errors

# Finding and fixing errors



- When the proactive approaches fail
  - unless you have very strong operational checks in place, and you can catch errors on a PR level
  - Helpful to have your data catalog update or DBT update (if the catalog depends on that) built into your deployment pipeline
  - Versioning is super helpful
    - Can even be git

# Closing thoughts

- These were just a couple questions through the proactive vs reactive lens
- Reactive has its place, **after the proactive** approach fails
- **Keep all the data** you can, but be conscious about **what you display**
- A lot depends on your organisation structure
  - Do Data Scientists write ETL jobs?
  - Are only Data Engineers responsible for operations?

# Thank You

**stemma**<sup>TM</sup>

from the creators  
of amundsen