coalesce

From coast to coast: Implementing dbt in the public sector

Jenna Jordan, City of Boston
Ian Rose, State of California
Laurie Merrell, Jarvus Innovations (Cal-ITP)

October 16-19, 2023

Meet today's speakers, working on public sector dbt projects

Jenna Jordan

Data Engineer

City of Boston Analytics Team

Ian Rose

Sr. Data Engineer

CA Office of Data and Innovation

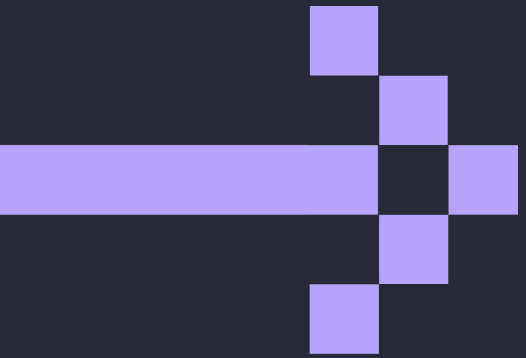
Laurie Merrell

Sr. Analytics Engineer

Jarvus Innovations

Agenda

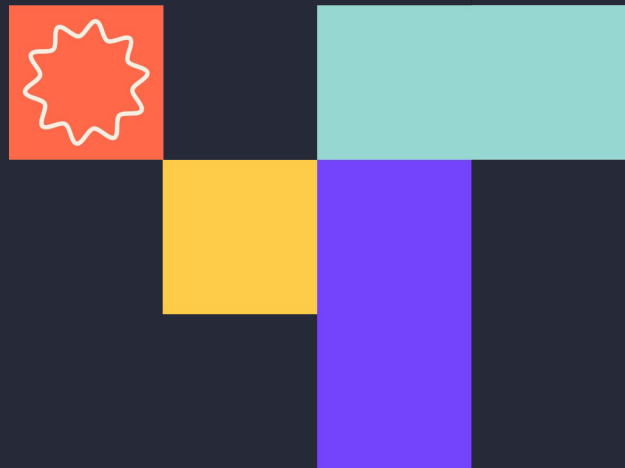
1. **Intros**
   To the speakers and case studies
2. **Public sector vs private sector data work**
   What makes government data work unique
3. **The case for dbt**
   Why now is the right time for government data workers to implement and use dbt
4. **Case Studies**
   Boston
   California (Hiring data & Public transportation)
5. **Lessons learned**
   Key takeaways to apply to your own projects
6. **Resources**
   Get connected to the community of dbt users in the public sector and available resources



If you are a public sector data worker, should you implement dbt? If so, what is the best implementation strategy for your team?

Introductions

To the speakers and case studies



City of Boston Analytics Team

City of Boston Analytics Team (DoIT)

- Founded in 2015; Currently 15 team members
- Analytics team works with departments on specific scoped projects
- 7 engineers (3 city employees + 4 contractors)
- Data engineers manage ETL pipelines + data warehouse, which the analysts use to produce reports & dashboards
- (Data Engineering) Tools:
  - Civis Platform
  - PostgreSQL + PostGIS
  - Python for custom ETL scripts
  - YAML for Civis workflows
  - SQL for data transformations
  - Great Expectations for DUTs
  - … and now dbt core!

California Office of Data and Innovation

- CalData is a division of the Office of Data and Innovation, a new State department as of this year!
- We act as internal consultants, researchers, and solutions architects for the State, helping improve government data operations
- Infrastructure built on Snowflake, dbt, PyData libraries, Airflow and AWS. But the State is large and we have to be flexible, so Azure, GCP, Oracle, etc are in the mix as well

Office *of* Data and Innovation

JARVUS INNOVATIONS

CAL ITP

Jarvus Innovations & Cal-ITP

- **Jarvus Innovations** is a tech strategy and engineering consultancy focused on frontline public services
- Managed by Caltrans, the **California Integrated Travel Project (Cal-ITP)** is a statewide initiative designed to unify transit in California through various data, payments, and standardization efforts
- Jarvus involved since 2021
- Data stack: Airflow, Google Cloud, dbt, Metabase

Public sector vs private sector data work

What makes government data work different

Processes can be more difficult

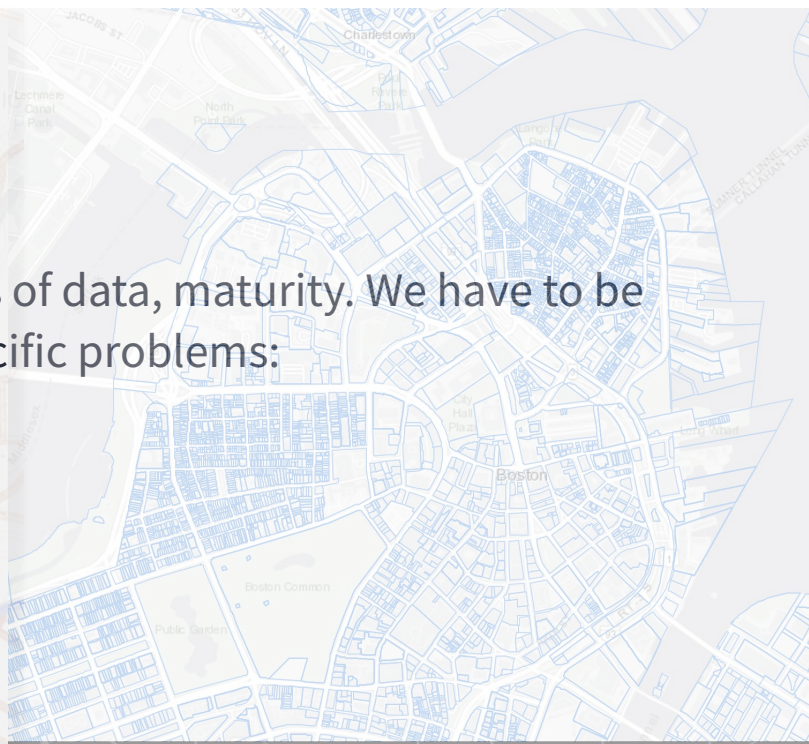
- Procurement can take months or years
- Purchase orders can be legally and logistically complex
- Budgets are limited and on a yearly cycle
- There is a culture of waterfall-style project management: building monolithic enterprise solutions and then going into maintenance mode
- There’s lots of siloed data: it can be difficult to get departments even within the same government to share data
- Engineering decisions are often downstream of data governance policies, which can be much more difficult to update

10

The data looks different

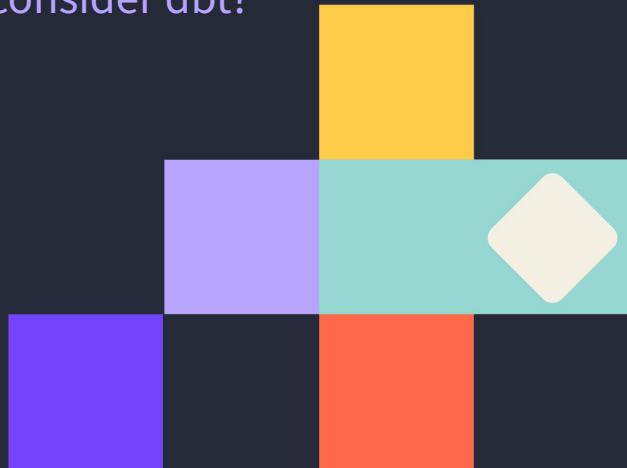
Extremely wide breadth of problems, types of data, maturity. We have to be flexible and ready to dive into domain-specific problems:

- Financial
- Geospatial
- Web performance
- Transportation
- Housing
- Human Resources
- Natural resources
- Climate
- … the list goes on!



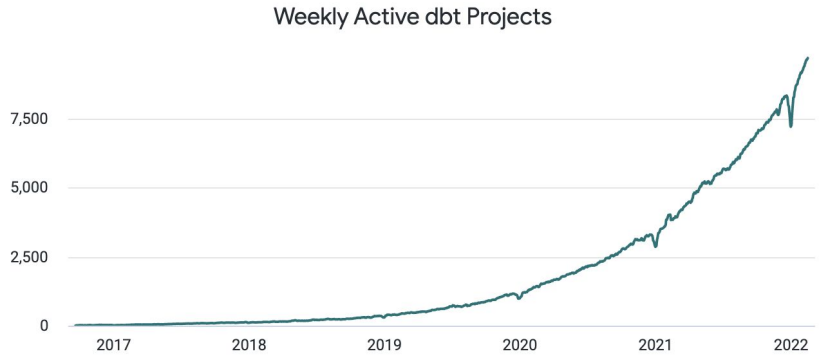
The case for dbt

Why is now a good time for government data workers to consider dbt?



dbt is a low-risk addition to your data stack

- Popular (over 30,000 companies)
- Mature (5+ years old)
- Free (open source Python package)
- Actively maintained/developed & documented by dbt Labs
- Foundational tool in the modern data stack
- Uses skills many organizations already have: SQL, YAML, Jinja



over 9,000 companies using dbt in production as of Feb 2022*

*over 30,000 companies using dbt in production as of today!
(As we learned in the keynote)

dbt may be a good fit for your public sector team if you…

Have a data warehouse or plan to have a central database for analytics

SQL as a common language for data… bonus if some users have git + command line experience

Are regularly ingesting new data and need it to flow through a transformation & testing pipeline

"dbt was built on the concept of taking the *best practices of software engineering* and blending them with analytics. I want to urge the political data community to *think about what best practices from Tech we can take and blend with our work* … I have seen firsthand what dbt can do for *small, under-resourced data teams* in politics."

- brittany bennett

https://www.brittanybennett.com/post/there-s-a-better-way-the-case-for-dbt-for-progressive-data-professionals

What does dbt provide to the public sector data worker?

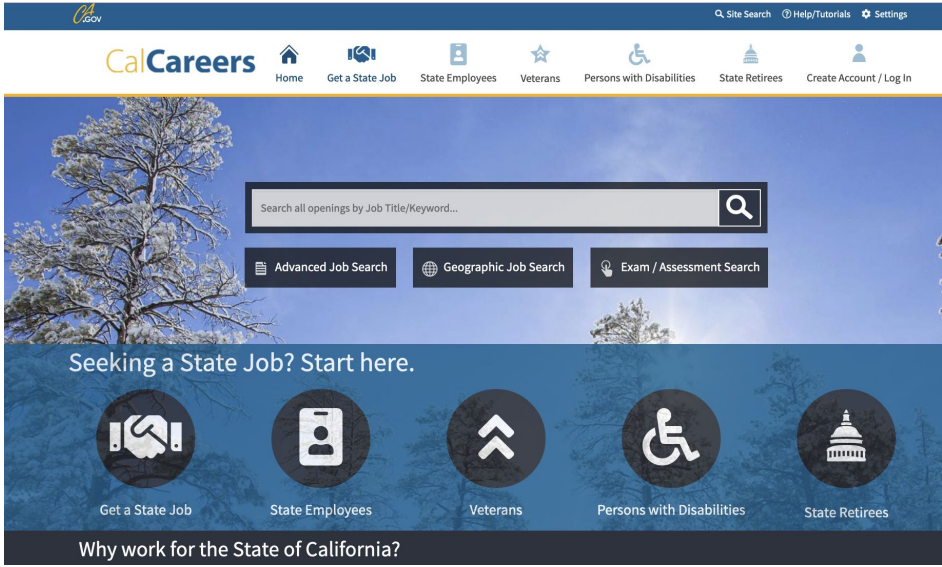
- Gets data transformations out of an analyst's head, and into version control!
- Encourages documentation of data models and automatically documents lineage
- Includes functionality for scheduling model transformations and ensuring data freshness
- Includes a testing framework for ensuring data quality/integrity
- Can work for both bottom-up and top-down data cultures
    - dbt Core: can start immediately without going through procurement, allows engineers to build a business use case
    - dbt Cloud: makes development easier/faster, enabling faster/broader upskilling & adoption

Case Studies

Getting down to the nitty-gritty

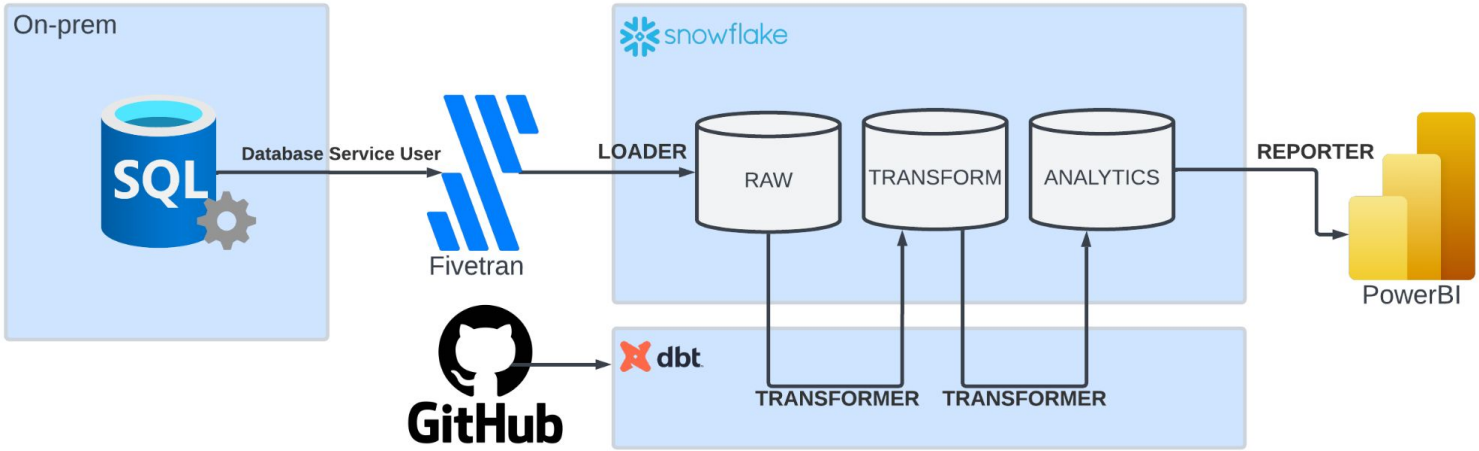
Case study: State of California Recruitment Data

- California is a huge state, with a quarter of a million employees and millions of job applications yearly
- California's recruitment analytics teams had challenges working with their data, including answering up-to-date questions about:
  - Hiring demographics
  - Strategies for job classifications and posting timelines
  - The effect of hiring campaigns
- Extremely familiar with their data, and with SQL
- Limited experience with version control, CI/CD, and scripting languages like Python

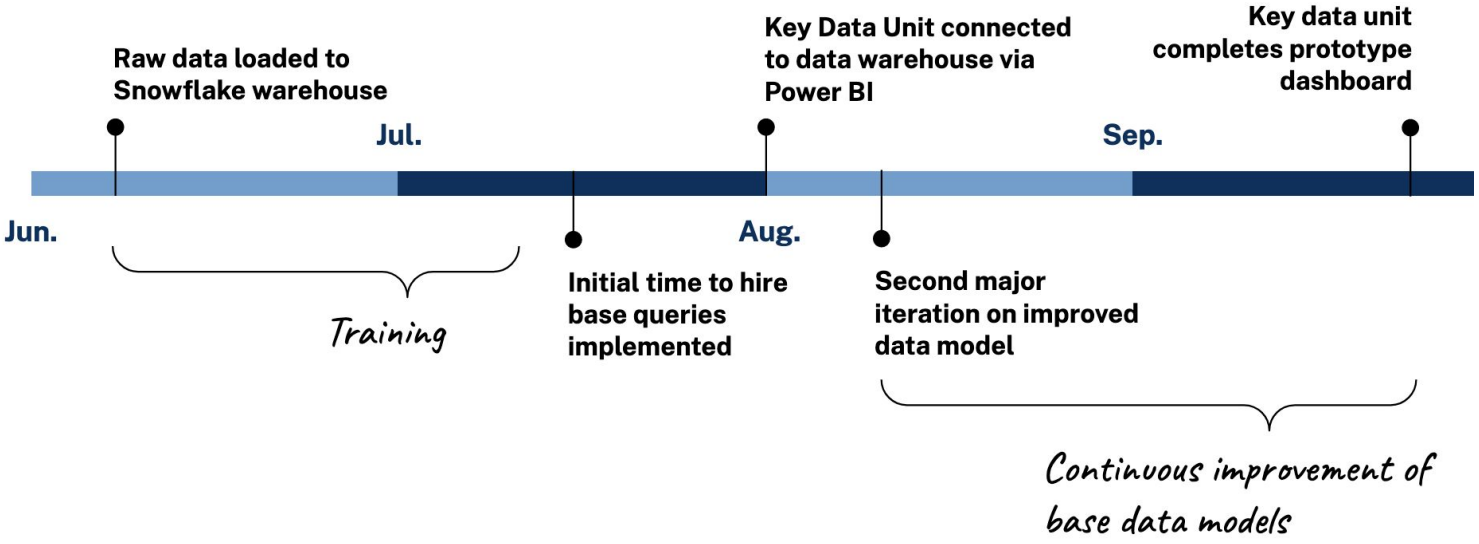


Case study: State of California Recruitment Data

Approach: string together Fivetran, dbt, Snowflake, and PowerBI for an entirely (well, mostly) SQL-based pipeline

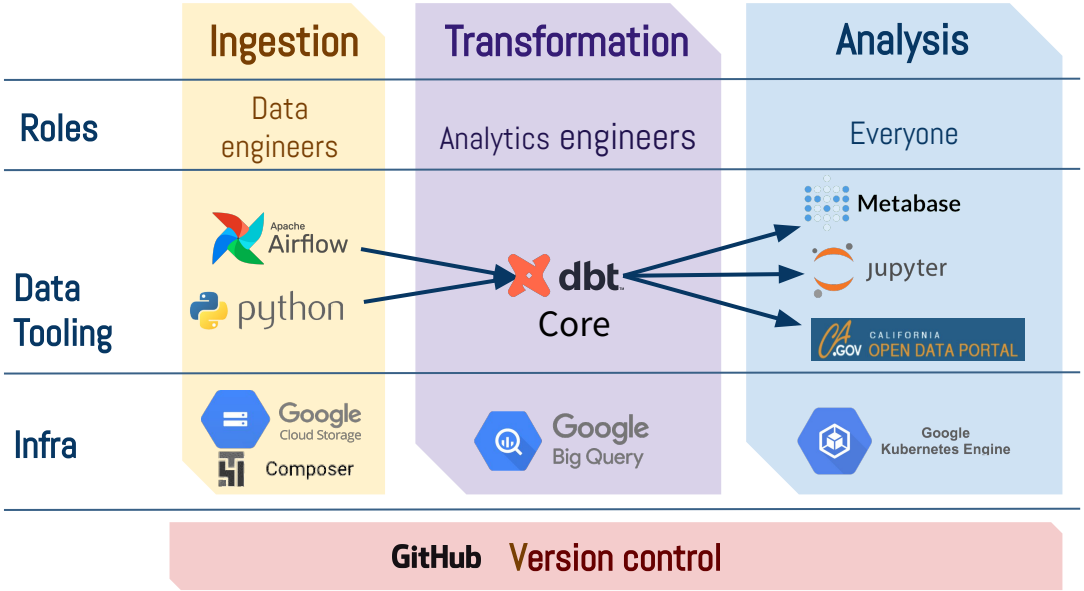


Case study: State of California Recruitment Data



Case study: Cal-ITP - Project context

- Inherited an MVP data pipeline using Airflow for SQL transformations
- Jarvus team:
  - 2 data engineers
  - 2 analytics engineers
- Supporting:
  - 5+ analysts
  - Customer success users
  - Transit agencies
- **Research** and **operational** users & use cases



Everything but dbt was in use before this engagement began

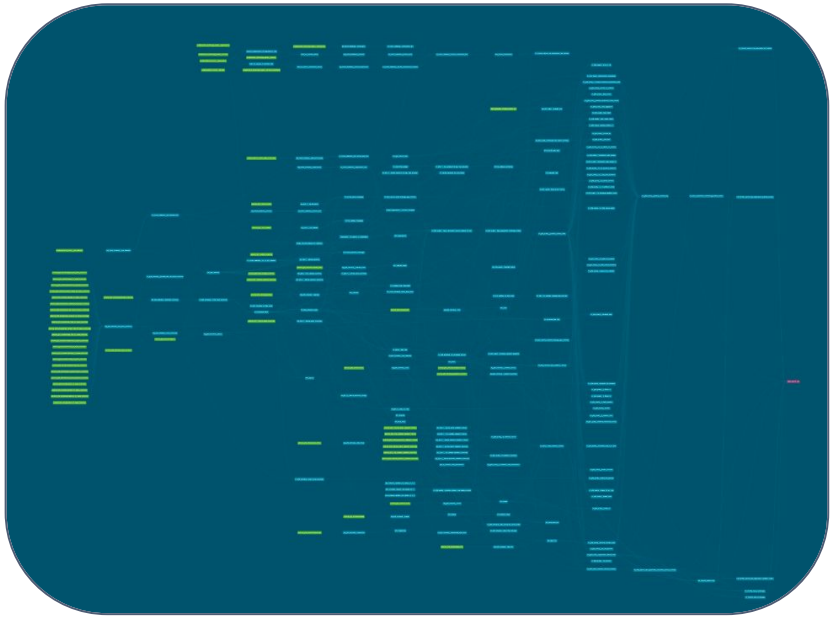
Case study: Cal-ITP - dbt implementation

APRIL 2022
Migrated data models from Airflow

MAY 2022
Proposed significant rewrite to deal with data ingest limitations

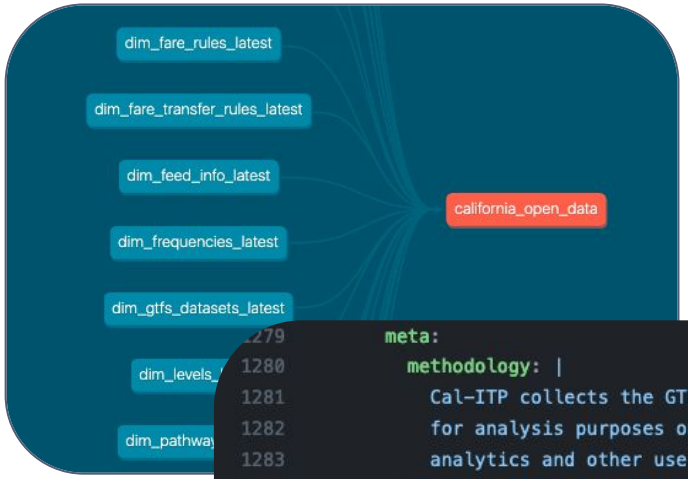
SEPTEMBER 2022
“V2” warehouse launched

TO PRESENT
Continue iterating: new models/features; deal with bugs; etc.



Open source repo:
github.com/cal-itp/data-infra

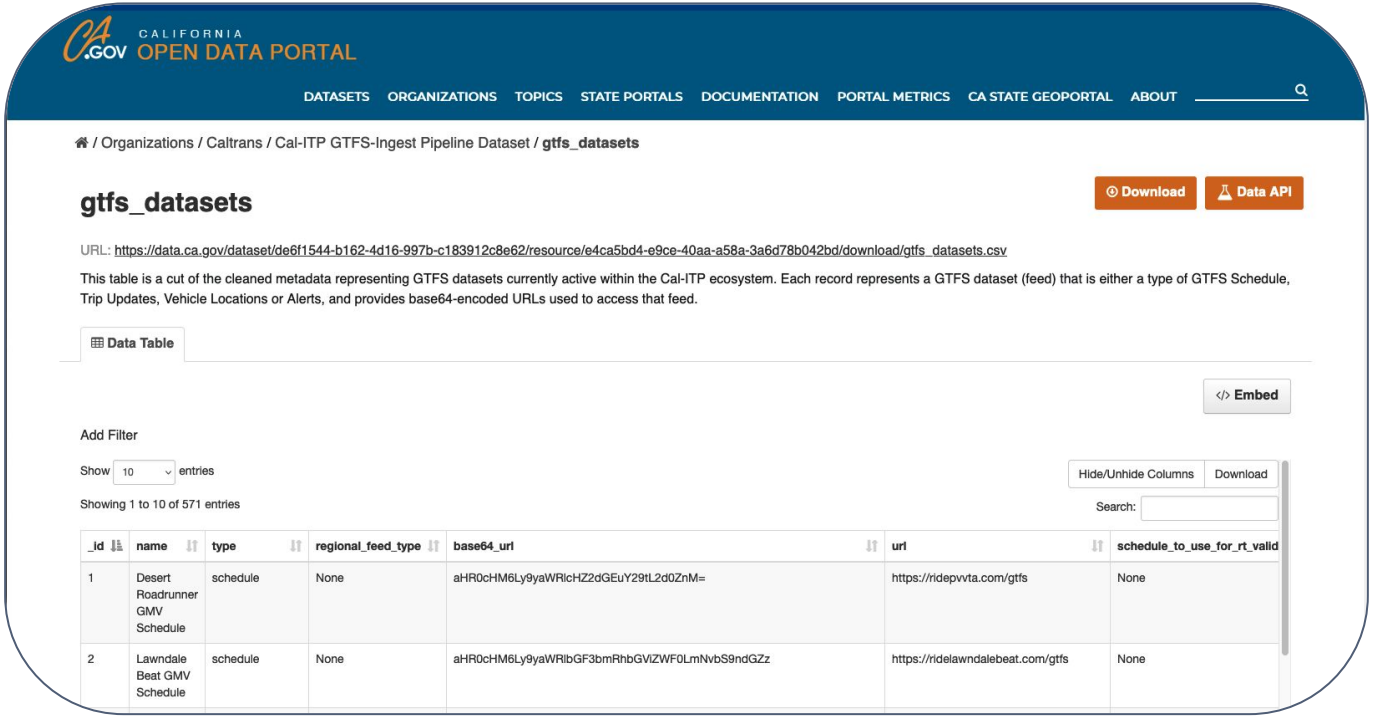
Case study: Cal-ITP - open data publishing with dbt



We use structured data from the dbt project (exposure, YAML config, manifest, etc.) to automate open data publishing to CKAN.

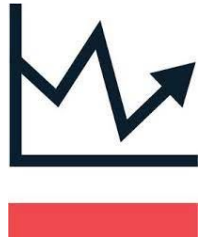
```
meta:
  methodology: |
    Cal-ITP collects the GTFS feeds from a statewide list every night and aggegrates it into a statewide table
    for analysis purposes only. Do not use for trip planner ingestion, rather is meant to be used for statewide
    analytics and other use cases. Note: These data may or may or may not have passed GTFS-Validation.
  coordinate_system_epsg: "4326"
  destinations:
    - type: ckan
      format: csv
      url: https://data.ca.gov
      resources:
        dim_agency_latest:
          id: c3828596-e796-4b3b-a146-ebeb09b3a4d2
          description: |
            Each row is a cleaned row from an agency.txt file.
            Definitions for the original GTFS fields are available at:
            https://gtfs.org/reference/static#agencytxt.
```

Case study: Cal-ITP - open data publishing with dbt



Open data at: https://data.ca.gov/dataset/cal-itp-gtfs-ingest-pipeline-dataset

Case study: Boston



State of the team

- Mature data analytics team with many *pre-existing pipelines* and data products (dashboards, open data, AGOL feature layers, etc)
- IT stack organized around *Civis*, an orchestration platform tailored to governments/nonprofits
- Existing orchestration platform + procurement constraints + team already using VS Code, git, command line = *dbt core* is a good option

dbt Implementation

- Needed a **redesigned set of schemas** first that worked best with dbt; allowed for dbt-oriented pipelines to be developed parallel with original pipelines
- dbt project really kicked off in sync with the **transition over to Power BI** (from tableau) - chance to reset on table dependencies
- Documentation site & lineage graph was a major selling point

Case study: Boston

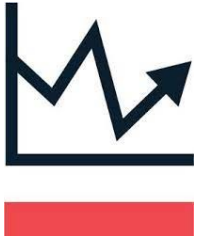
Each task executes in an ephemeral docker container

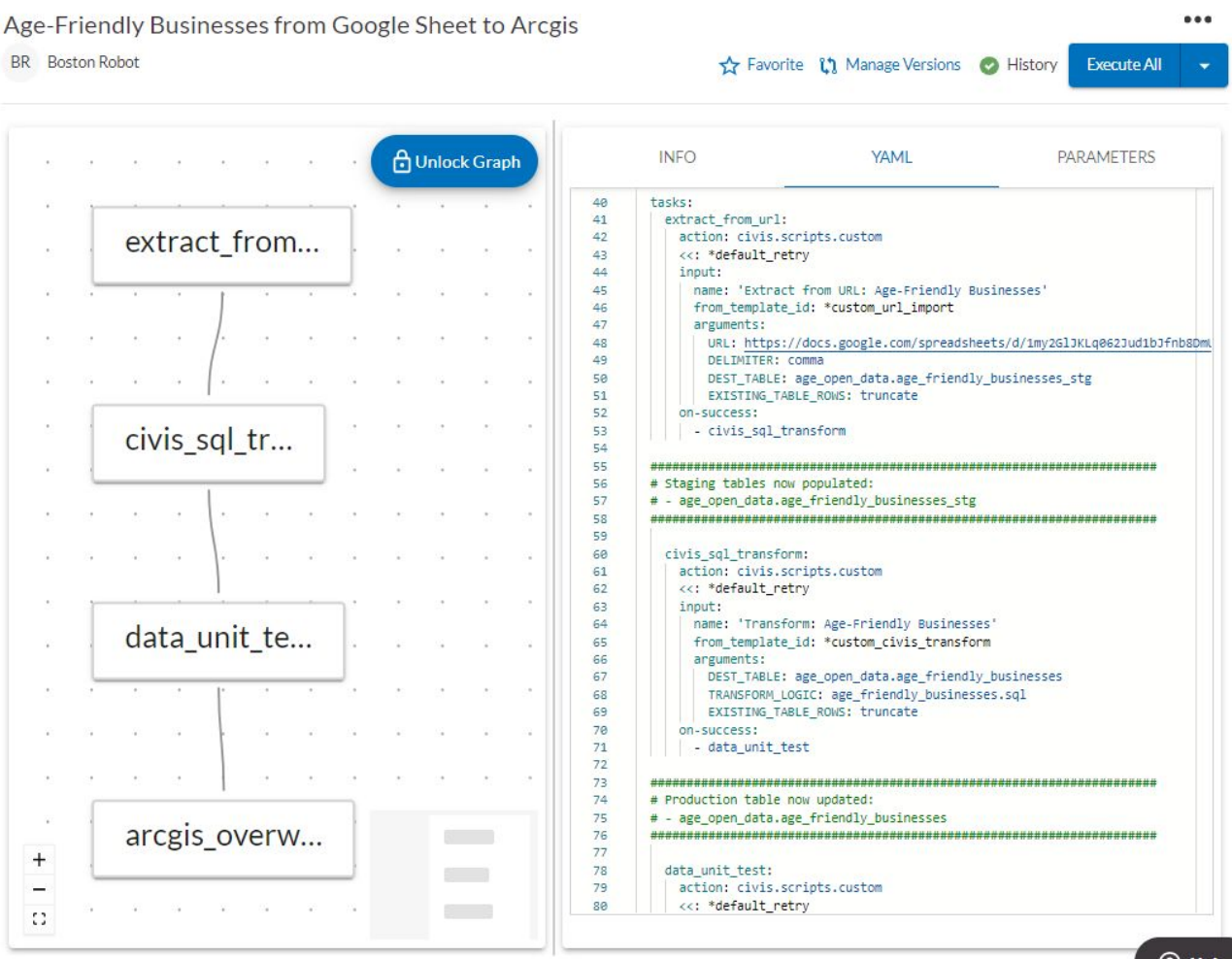
4 mins

5 mins
Execute SQL: < 1 min

1 min

4 mins





Case study: Boston

Some of the original set of schemas…

ALL SCHEMAS Name: A-Z

- age_internal_data
- age_open_data
- agilepoint_internal_data
- analytics_internal_data
- analytics_open_data
- analytics_restricted_data
- archives_internal_data
- archives_open_data
- arts_internal_data
- arts_open_data
- assessing_internal_data
- assessing_open_data
- audit_internal_data
- bais_internal_data
- bais_restricted_data
- bcyf_internal_data
- bcyf_open_data
- bfd_internal_data
- boundaries_open_data
- bpda_internal_data
- bpda_open_data
- bpd_internal_data
- bpd_open_data
- bphc_internal_data
- bpl_internal_data
- bpl_open_data
- bps_internal_data
- btd_internal_data
- btd_open_data
- disabilities_internal_data
- disabilities_open_data
- dnd_internal_data
- dnd_open_data
- doit_internal_data
- doit_open_data
- egis_internal_data
- elections_open_data
- env_internal_data
- env_open_data
- food_internal_data
- food_open_data
- food_restricted_data
- hansen_internal_data
- hansen_open_data
- hcm_internal_data
- hcm_restricted_data
- hhs_open_data
- hrc_internal_data

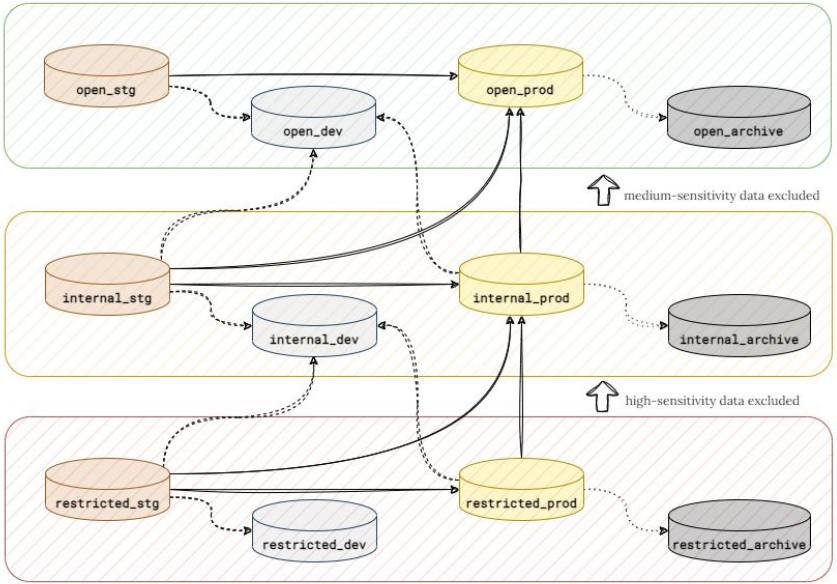
ANALYTICS DATA WAREHOUSE
NEW SCHEMA DESIGN

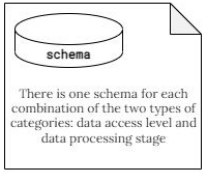
DATA ACCESS LEVEL

OPEN
- low-sensitivity data
- already on public record
- otherwise available to the public
- appropriate for unrestricted internal use

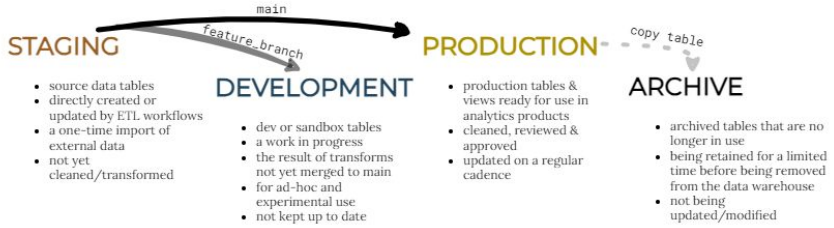
INTERNAL
- medium-sensitivity data and the default classification
- not already publicly available
- may include PII, especially in combo with other data
- should be restricted to the team that directly needs it

RESTRICTED
- high-sensitivity data
- has legal use restrictions
- contains sensitive unredacted PII that could cause harm
- requires strict control of access



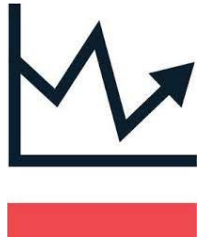


There is one schema for each combination of the two types of categories: data access level and data processing stage



DATA PROCESSING STAGE

Case study: Boston



Case study: Boston

MARCH 2023
Proposed, workshopped, and then created the new set of schemas

MAY 2023
dbt project repo setup finished, example models added

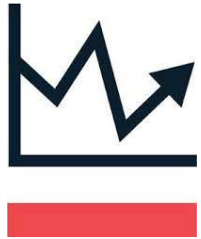
JUNE 2023
2 more engineers onboarded & start contributing to project;
Focus on adding core data sources (Hansen, 311, EGIS, etc)

JULY 2023
More engineers onboarded; focus on building out all dependencies for high priority PowerBI dashboards

SEPTEMBER 2023
Dependencies for all high priority PowerBI dashboards completed;
docs site is regularly updated & has Boston branding;
dbt workflows in production

OCTOBER 2023
Coalesce!

Case study: Boston



Goal	Pre-dbt Pain Point	dbt Value-Add
Data Catalog	• Some data sources documented, in many locations & formats	• Automatically generated & updated data catalog
Data Governance	• Data lineage, ownership, downstream use, and freshness are unclear and not documented	• Data lineage automatically documented & visualized • Data ownership explicitly & centrally documented
Change Enablement	• Data warehouse is a black box for those not on Analytics team	• Automatic dependency graphs, including Exposures (external dependencies) on data catalog site
Faster Outcomes	• Always seeking continuous improvement	• Packages & macros enable DRY code • Declarative (dbt handles execution)
Data Quality	• Test failures not accessible/transparent	• Easier to add & create tests • Test failures recorded

Lessons learned

Key takeaways to apply to your own projects

Lessons learned

- Be flexible & meet people where they are
  - Culture shifts take time
  - Analyst adoption requires care
  - Building consensus is key
- Balance incremental adoption vs. leveraging synergy
  - Quick, cheap wins build buy-in
  - But some changes are easier when done together
- Iteration is good: dbt allows you to iterate & lessens cost of “mistakes”
  - This is not a “one and done” installation
- Adapt best practices to your team’s unique environment
  - Don’t be afraid to “break the rules”

Help us build a public sector dbt user community!

- GitHub repos:
  - github.com/jenna-jordan/dbt-public-sector-resources
  - Cal-ITP data infrastructure
  - CalData data infrastructure
  - CalData project template
  - Boston dbt project skeleton
- Continue conversations in the **#industry-public-sector** channel in dbt Slack

Thank you

This presentation recording will be sent out shortly