

# Mastering Multitenant Orchestration with dbt and Dagster

Life of Data Engineers does not have to be that hard

Andrea Montes

DBT Bogotá Meetup

2024-06

## Audience questions

### Problem

#### We need to re-design

- Requirements

- New architecture

- Stage: Extract

- Stage: Load

- Stage: Transformation

- Stage: Transformation - Leveraging DBT for multitenancy

- Stage: Transformation - but for +250 clients?

### Takeaways

# Audience questions

1. Airflow familiarity?
2. Crazy tools difficult to debug?

# What is needed?

- ▶ Daily updates to client dashboards. +250 different clients
- ▶ Custom reports per client
- ▶ Product and business questions

# Legacy product(s) - Data Warehouse

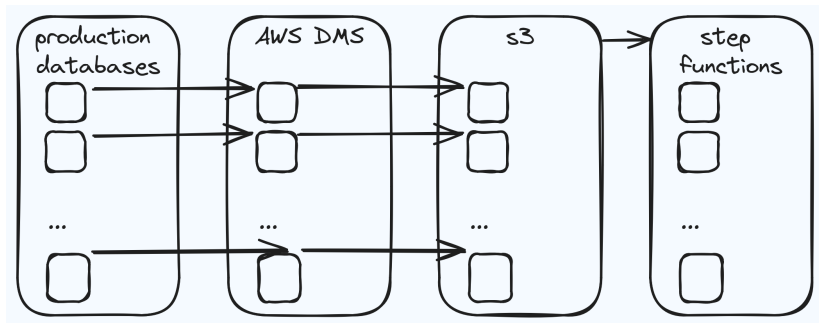


Figure: Legacy DWH product

- ▶ Daily refresh
- ▶ Failed 2 or 4 times a week for heavy clients 🙄

## Legacy product(s) - Reporting

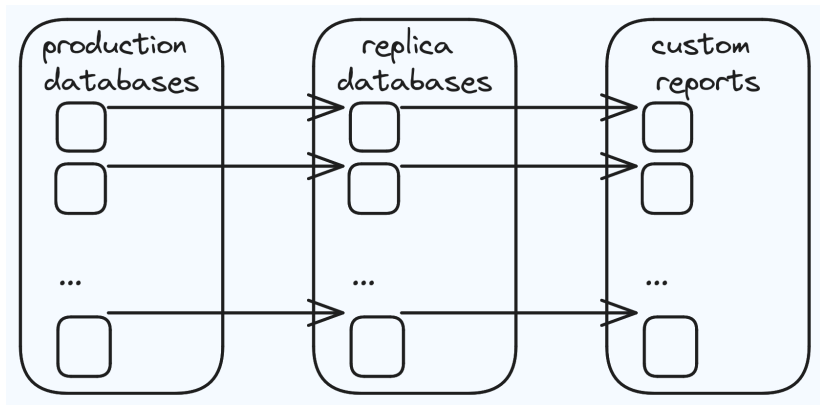


Figure: Legacy reporting product

- ▶ Replicated once a week 😞
- ▶ Raw sql
- ▶ Stored procedures

# We need to re-design

# Different users, different requirements

## Dashboards

Clients need their dashboards updated to know events statuses

## Business questions

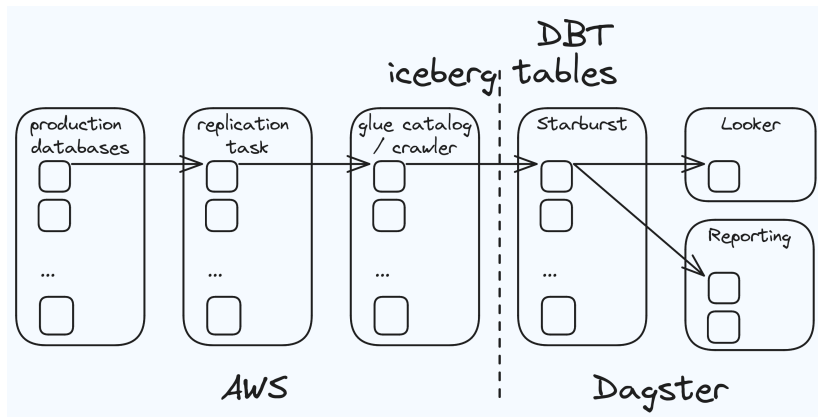
What are the most used features?, how virtual vs in-person events attendance has changed after pandemic 🧐?

## Client questions

How long is taking a candidate to become an applicant? Do I have a diverse pipeline?



# New architecture



**Figure:** New architecture to accomplish reporting and DWH requirements

# DMS

- ▶ Cheap...
- ▶ Not reliable as we would like
- ▶ Trade-off

# Glue catalog and Crawler

- ▶ Easy enough to implement
- ▶ Our compute engine support it

# Starburst - DBT



- ▶ Starburst is a vendor option for Trino
- ▶ Cheaper than snowflake
- ▶ Trino has a good community



- ▶ SQL ANSI
- ▶ It's a new product, random changes
- ▶ Compute throttling
- ▶ We were their first large user

# DBT and multiple clients

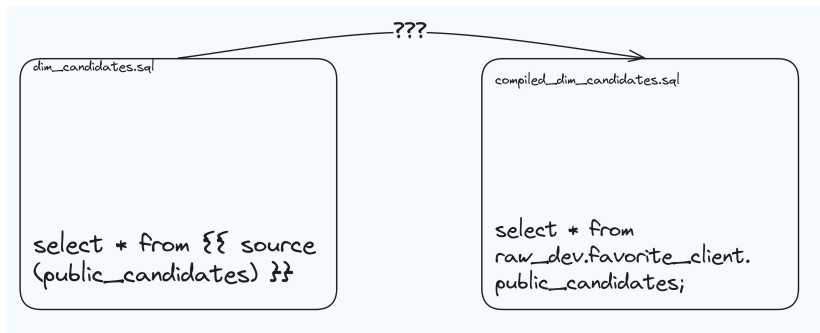


Figure: How to do multitenancy in dbt

# DBT and multiple clients

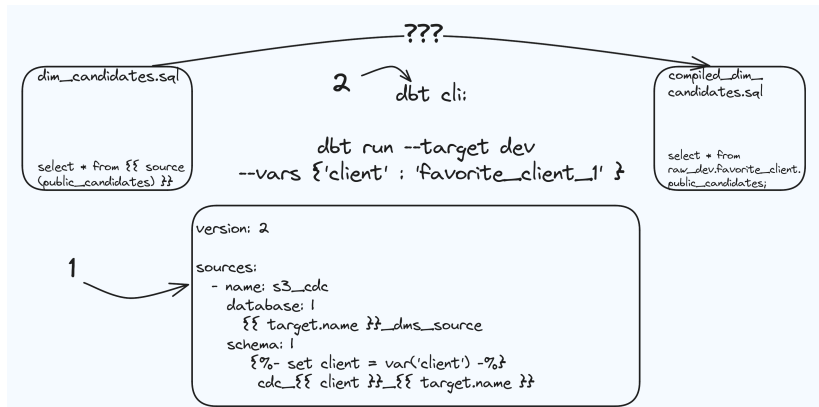
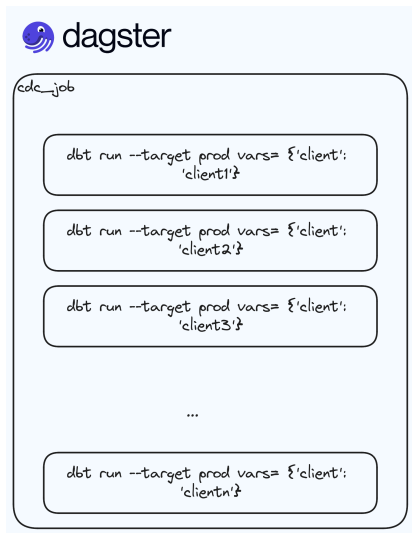


Figure: Multitenancy sorted out using dbt cli and variables

# Orchestration tool: Dagster



- ▶ We run `cdc_job` daily starting at 2 am CST
- ▶ Each job client partition takes about 7 mins
- ▶ Job finishes before 7 am CST

Figure: Dagster: How do we run 250 dbt commands?

# Takeaways


- ▶ Not getting used to what is poorly done
- ▶ Don't fall in love with a technology product




# Who am I?



Figure: Andrea Montes

 Andrea Montes - Senior Data Engineer

 mamontesp