Welcome

Chicago dbt Meetup

Host: Analytics8

Join dbt community Slack
getdbt.com/community
#local-chicago

MC: John Barcheski

Manager, Technology and Strategy & dbt Practice Leader

Analytics8

@John Barcheski

jbarcheski@analytics8.com

Linkedin.com/in/barcheski

Transform your
organization
with data.

Agenda

- 5:30 – Doors Open and Check-in
- 6:00 – Speaker Sessions

  -Tyler Rouze, Analytics8

  -Jack Blandin, Fi
- 7:00 – 9:00 – Reception

Register for Coalesce 2023

The Analytics Engineering Conference
October 16-20

Coalesce Online **Coalesce London** **Coalesce San Diego** **Coalesce Sydney** **Coalesce Berlin**

for 15% off, use code **dbtmeetup15**

Coalesce Online
October 16-20, Online
FREE
The main experience you know and love.
Visit **coalesce.getdbt.com/register**

Coalesce San Diego
October 16-20, San Diego
Take advantage of Early Bird rates for the week long event in San Diego until February 28th. Visit
coalesce.getdbt.com/register

Coalesce in EMEA
We will be hosting one-day Coalesce events in London, Sydney, and Berlin as well. Registration will open for those locations in early 2023.

dbt

Speaking Today

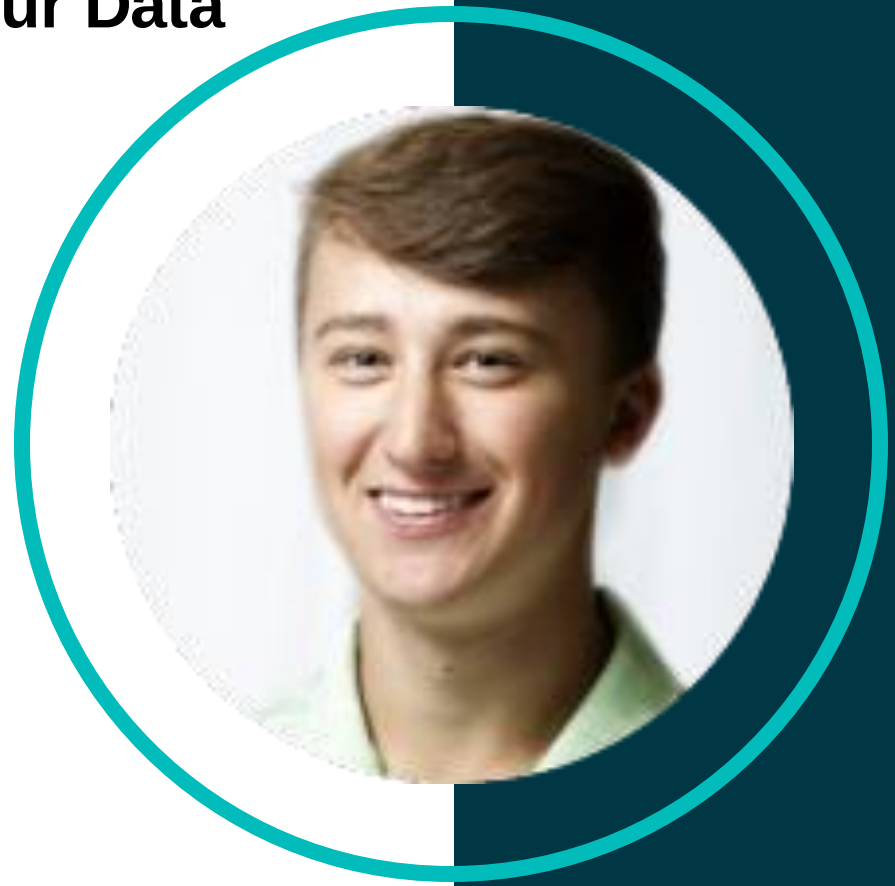
Tyler Rouze

Analytics8

Jack Blandin

Fi

Lessons in Scaling Your Data Platform with dbt



Tyler Rouze

Senior Consultant, Analytics8

Linkedin.com/in/tylerrouze/

trouze@analytics8.com

tylerrouze.com

~$ whoami

- Working alongside Roche Pharma Commercial on large scale implementation of dbt
- Tens of thousands of models
- 100s of developers across 41 countries and counting
- Most contributed repositories in a 100k employee company

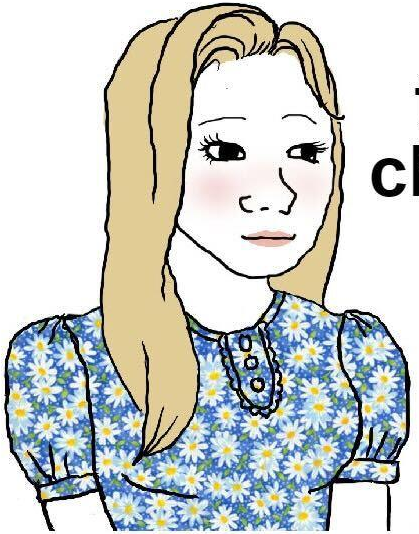
Analytics8 Roche

WEBINAR
How dbt Helps Our Clients Remove Data Engineering Bottlenecks
John Barcheski
Data Engineering Consultant
Erin Barnier
Data Engineering Consultant
Tyler Rouze
Data Engineering Consultant
Ryan McGibony
Partner Manager @dbt Labs
Jordan Veal
Data Engineering Consultant
dbt
Analytics8

Webinar

Power BI
Gebruikersgroep
hillstar
business analytics
GUY in a CUBE
GUY in a CUBE

Analytics8



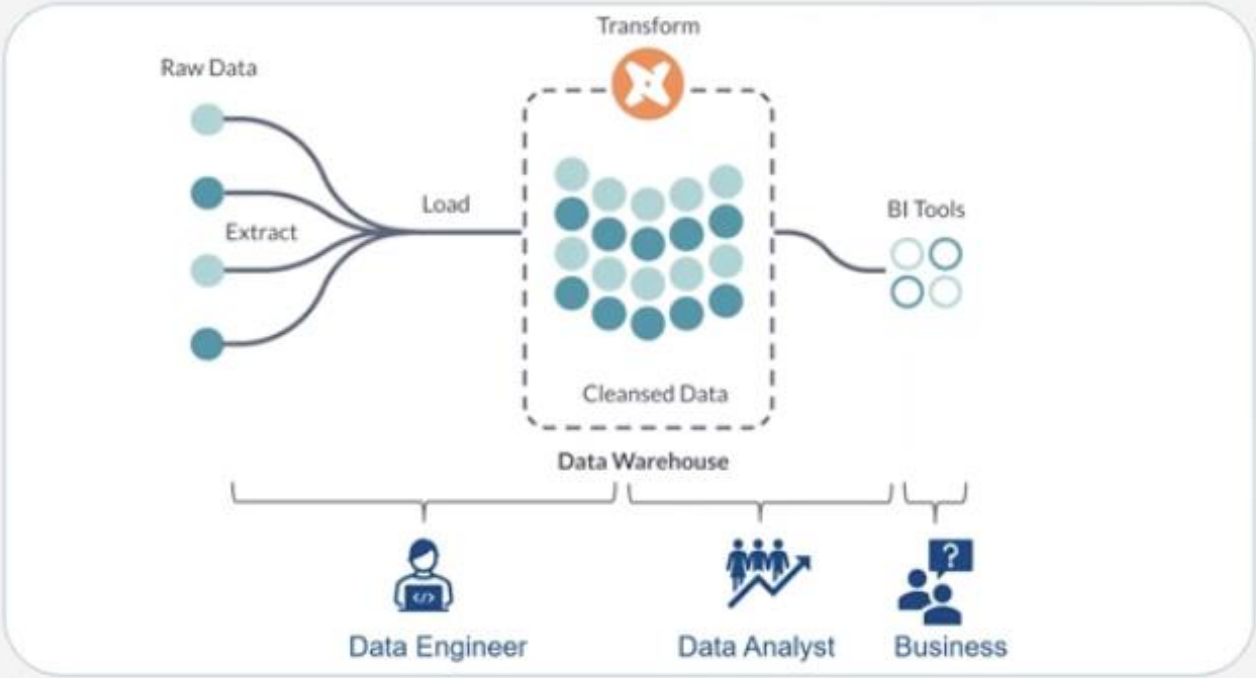
thank you for
changing my life

dbt

I'm literally just a Jinja templater that runs your SQL in the right order

CENTRALIZED DOCUMENTATION & LINEAGE

Bring your data analysts, BI developers, and SQL developers into your data pipeline development to build trust with stakeholders



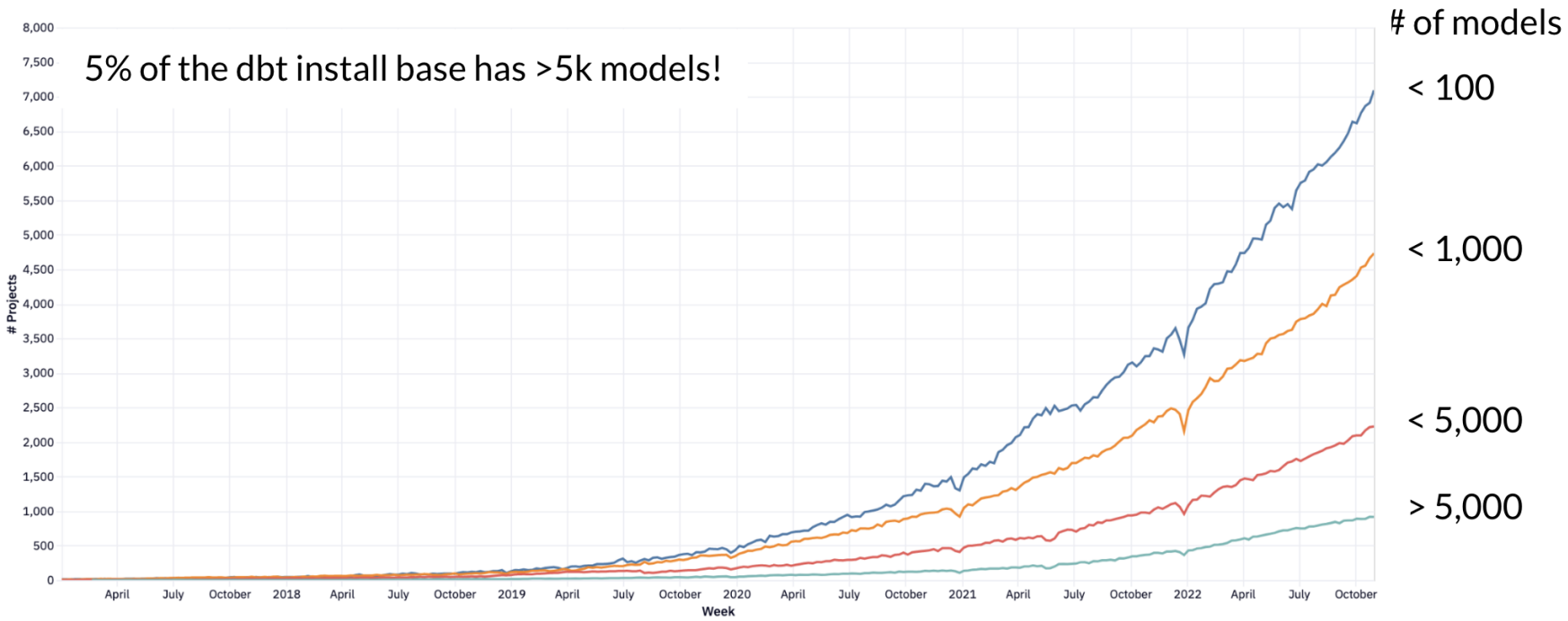
Goals of scaling our data platform with dbt

Ultimately, increase the value of your company's data

Enable heightened speed of development over IT, more consistency over Business

Meet your talent where it's at and support their involvement on your data platform, keeping guardrails in place

dbt is the linchpin in keeping everything-as-code on your data platform


5% of the dbt install base has >5k models!
# of models
< 100
< 1,000
< 5,000
> 5,000
Projects
8,000
7,500
7,000
6,500
6,000
5,500
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
April
July
October
2018
April
July
October
2019
April
July
October
2020
April
July
October
2021
April
July
October
2022
April
July
October
Week

Where your organization is coming from influences your project architecture

People coming from IT with different skillsets (Talend, Informatica, Oracle)

Limited competencies in modern data tooling

Build a new data platform on modern tooling that supports numerous geographies and their analytics needs

Own the architecture, but support and enable horizontal scaling of the platform

dbt Project Federation

Analytics8

Federated dbt Project Architecture supports decentralized ownership across skillsets.

Analytics8

Project federation supports different workflows and code ownership

Allows your dbt projects to scale horizontally to touch further parts of the organization

Can help make core logic reusable across the organization

Fosters collaboration and knowledge sharing across projects

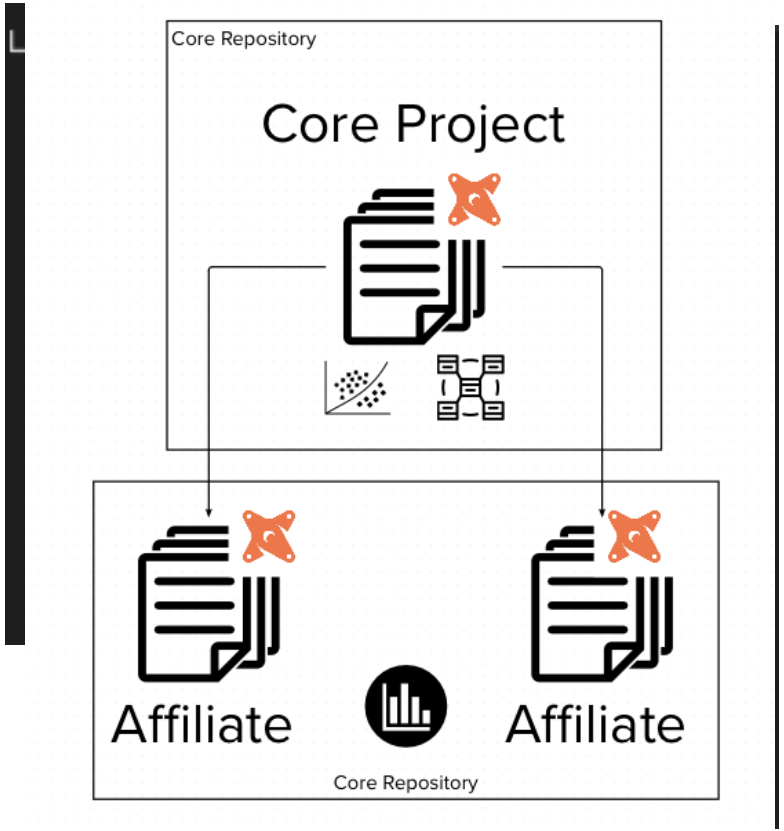
Resource efficient

Federated dbt Project Architecture supports decentralized ownership across skillsets.

Analytics8

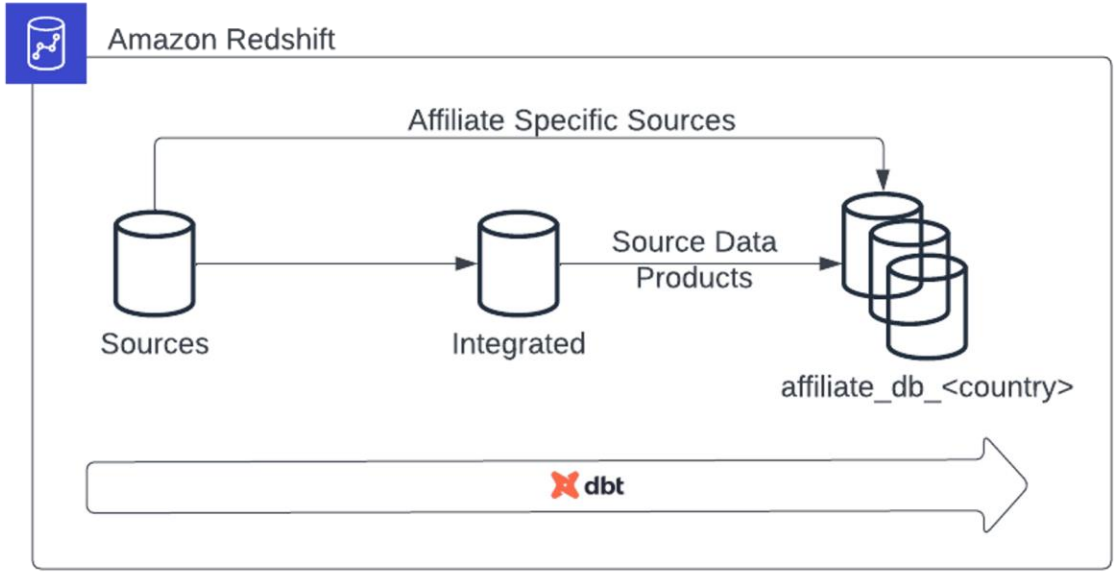
Core project Supports Data Modeling & technical dbt modeling (ML, Embedded Apps)

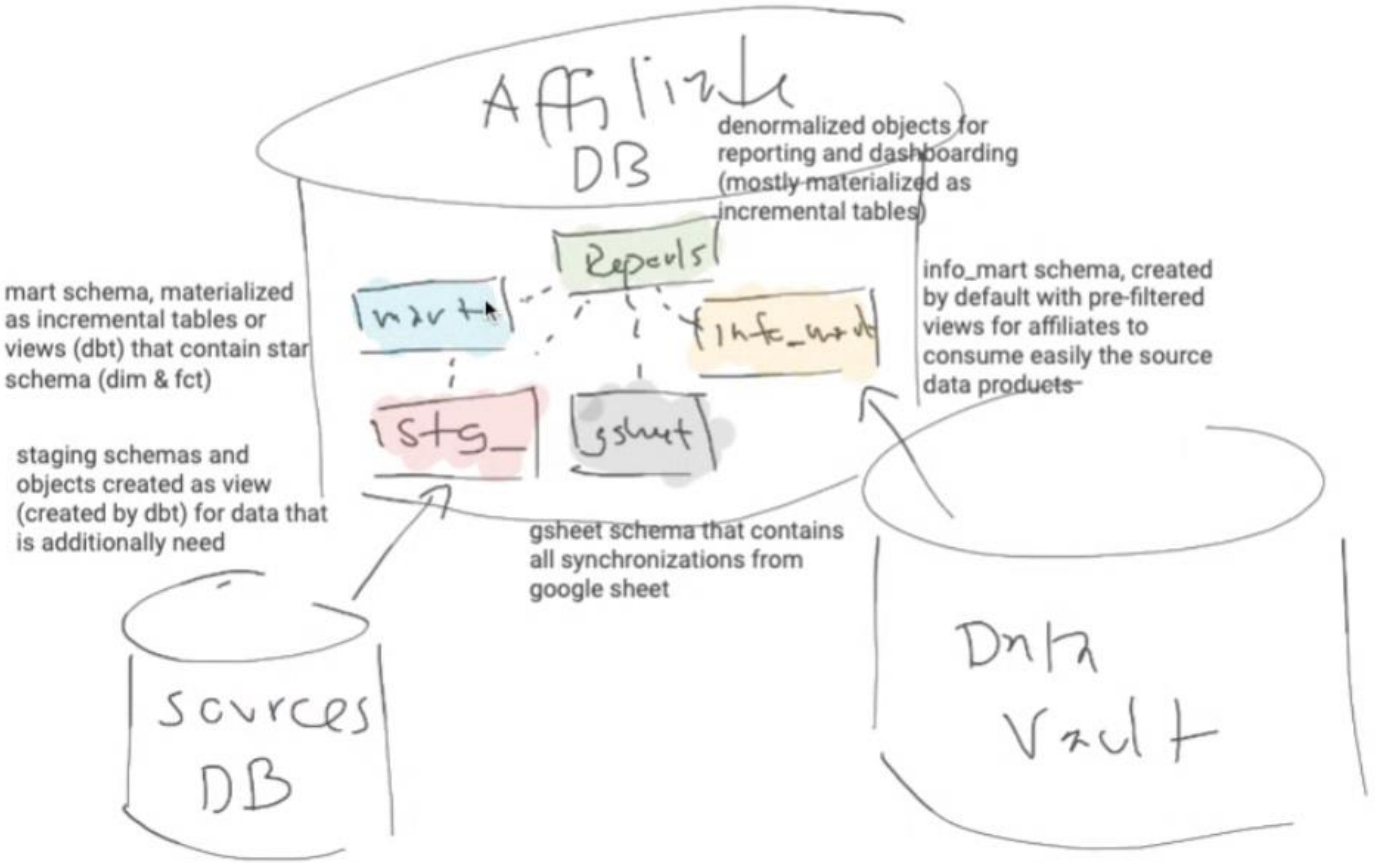
“Affiliate” projects ensure decentralized ownership while still leveraging ”Core” products & best practices



Single Cluster, multi-database supports flexibility at scale

Analytics8




Affiliate DB
denormalized objects for reporting and dashboarding (mostly materialized as incremental tables)
Reports
mart
info_mart
mart schema, materialized as incremental tables or views (dbt) that contain star schema (dim & fct)
info_mart schema, created by default with pre-filtered views for affiliates to consume easily the source data products
stg_
gsheet
staging schemas and objects created as view (created by dbt) for data that is additionally need
gsheet schema that contains all synchronizations from google sheet
Sources DB
Data Vault

Common elements can be shared across a monorepository

We want to maximize reusability while decentralizing ownership across a federation of projects

```
packages:
  - local: ../../common/dbt
  - local: ../../../../common/dbt
```

Common elements can be shared across a monorepository

```
{{ config(schema="info_mart",tags="info_mart",materialized='view',bind=false) }}
SELECT
 *
FROM
    {{source('info_mart','dim_address')}}
    WHERE 1=1
      AND row_level_security_country_code in {{ var("country_code") }}
```

dbt-core & dbt Cloud in Conjunction

Analytics8

Hadrien 5:27 AM

Hello fellow dbt user! ✖️💪

In my company we have a development bottleneck. Since our team grew a lot recently we have issues deploying changes continuously 😕
Let me explain our development process, it should also be noted that we are using Redshift as the DWH and Gitlab CI/CD to manage releases.

Radek Tomsej 8:48 AM

Hi all 👋 I really ❤️ `dbt`! However, I found that our dbt project grew significantly, it became really difficult to maintain the quality of the models. People often forget to add or delete a column, description or edit a test in a schema file. It is practically impossible to catch all these problems during the review process. In addition, with a large number of objects in dbt, running models is slow and expensive, so it's a good idea to find as many errors as possible before the changes are committed and pushed to branch.

osmang 10:48 AM

👋 Hi everyone,
We are a couple of data / analytics engineers trying to build some automation and tools around the development process. ⚙️🏗️
There are several segments of our processes that we feel are slow and time consuming, and we are curious about whether others have similar experiences. ⏳

Pawel Kapuscinski 8:10 AM

hey team. I'm looking for good resources on git and software development process for non technical people who just try to understand better what's happening in their companies. Any recommendations?

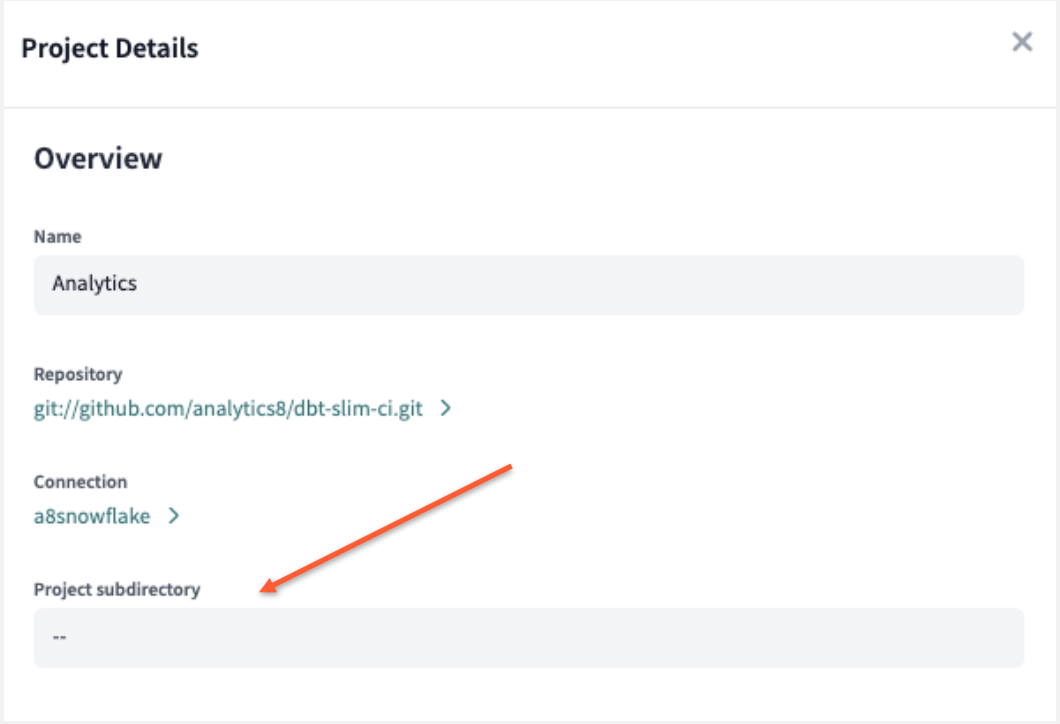
Meet your team where there at with separate development setups

Core Engineer

Support source-code editor of choice

Affiliate Analyst

Make use of dbt Cloud to onboard into platform development


Project Details
Overview
Name
Analytics
Repository
git://github.com/analytics8/dbt-slim-ci.git
Connection
a8snowflake
Project subdirectory
--

People & Development Workflow

Analytics8

How do we structure our people to get the most out of our project federation

Matrix organization

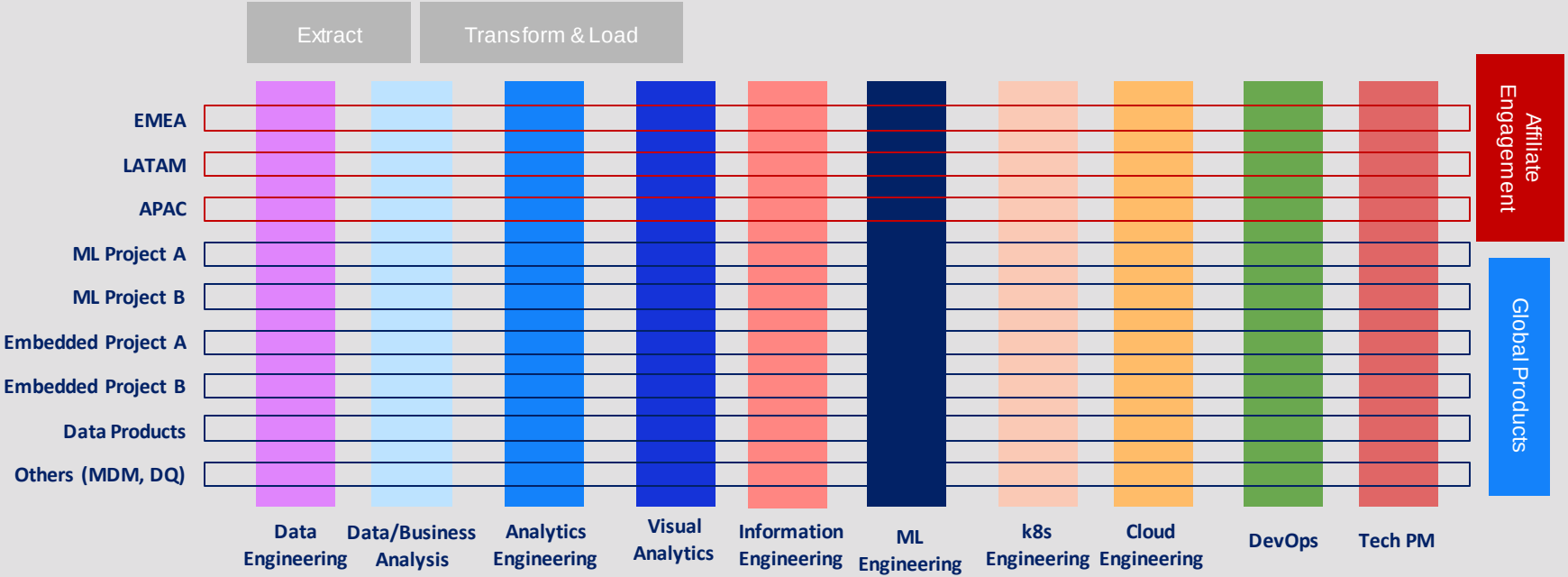
Share technical competencies across the organization (CI, Kubernetes, Orchestration)

Engineering leadership over regions of federated projects

Engineering Organization

Head of Engineering

Lead Engineer



CAPABILITIES

Principles

Operate on a regular sprint schedule across projects

Rotate operations team responsibilities

In federated monorepo, utilize code ownership scoping with leadership across multiple projects

Everyone should be familiar with your Git Solution (Gitlab, Github)

Keep documentation close

Job Scheduling

How should I run my jobs in this architecture?

It's okay to support scheduling in multiple ways, both dbt Cloud jobs and a high code option

Regardless of how you run jobs, store configurations as code so job templates can be reused, redeployment is easy, and refactoring doesn't have to happen in a UI.

Leverage CI job templates to avoid repeating pipeline definitions

```yaml
dbt_run:
  extends:
    - .matrix
  image: ghcr.io/dbt-labs/dbt-redshift:1.5.0
  script:
    - cd ${DBT_PATH}
    - dbt deps
    - dbt build --profiles-dir . -t $TARGET
```

```yaml
.matrix:
  parallel:
    matrix:
      - NAME: PROJECT-1
        DBT_PATH: projects/path/to/project/dbt
        OTHER_VAR: xyz
      - NAME: PROJECT-2
        DBT_PATH: projects/path/to/project/dbt
        OTHER_VAR: abc
```

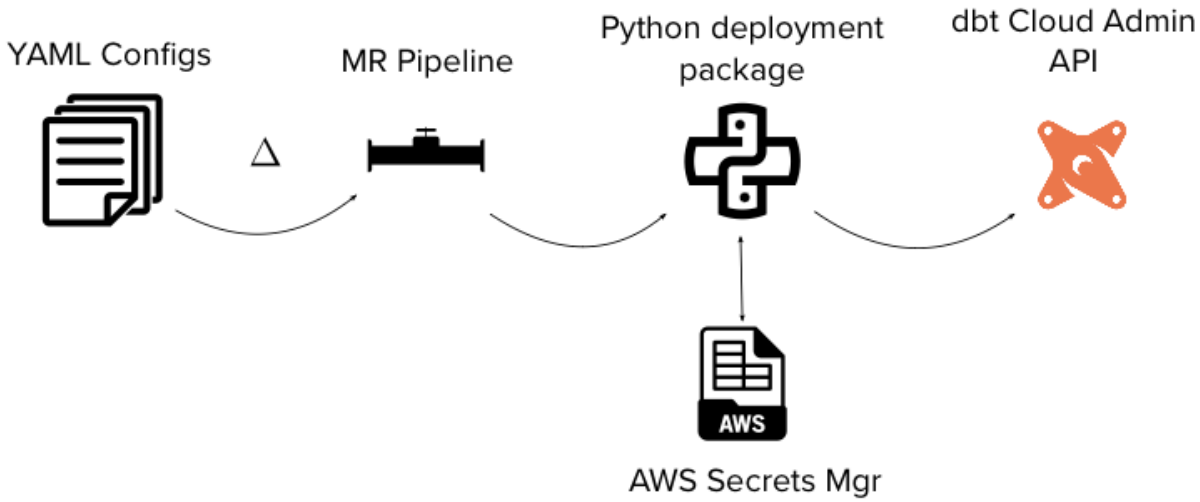
If you’re using an orchestrator that runs granular jobs, consider job startup latency

1. Await Container Creation

2. Pull Git Repository

3. Install dependencies `dbt deps`

Leverage dbt Cloud Administrative API to deploy projects, environments, and jobs



Wrap Up

Why do this?

Increase speed of development while maximizing reusability (and thus minimize costs) while supporting horizontally scaling your company's data

Build trust with the business and deliver on their use cases

Increase the value of your company's data

What's Next?

dbt v1.5 & looking forward to v1.6

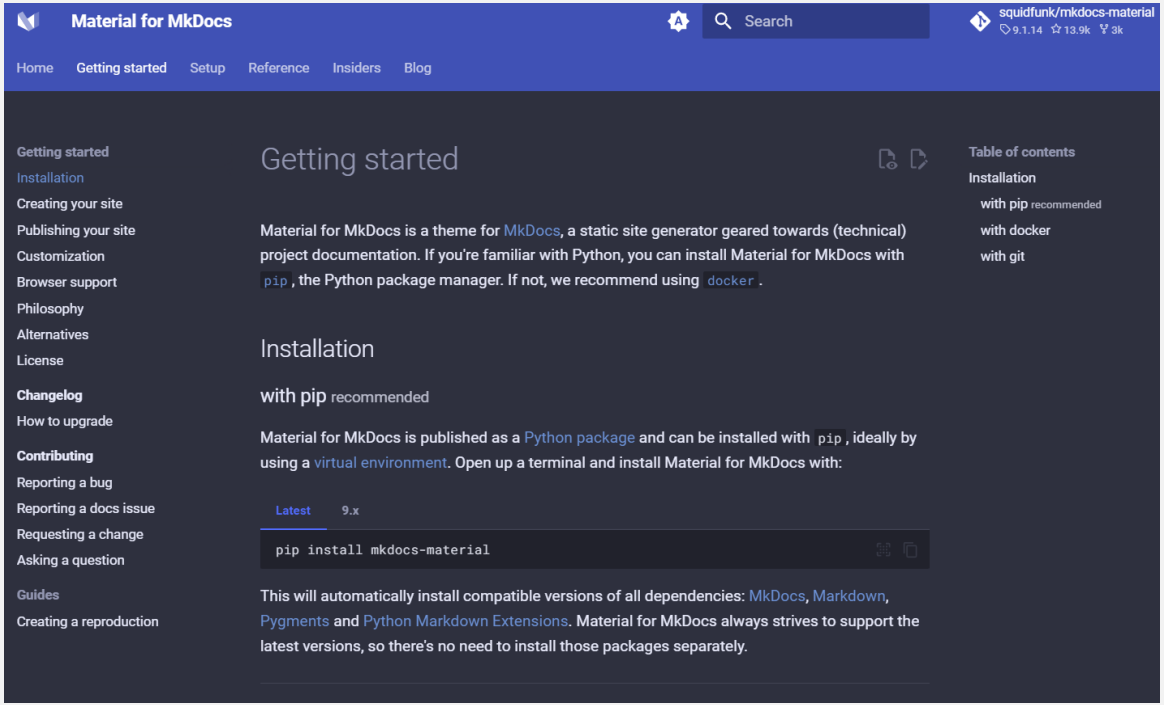
Model Contracting, Versioning, and Access

Cross-project `refs` with state handling

Transform your
organization
with data.

Appendix

Documentation



With a single environment, utilize proxy views to scale development while preventing unnecessary data copying

upstream_prod

Created by LewisDavies

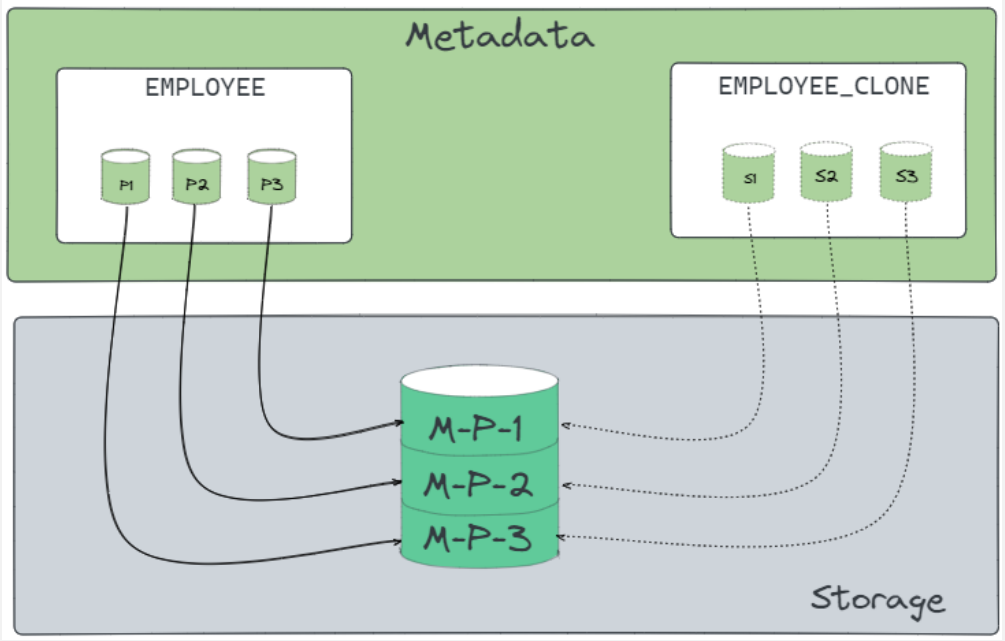
LewisDavies/upstream-prod 17

Installation

dbt version required: `>=1.1.0, <2.0.0`

Include the following in your `packages.yml` file:

```
packages:
  - package: LewisDavies/upstream_prod
    version: 0.4.0
```

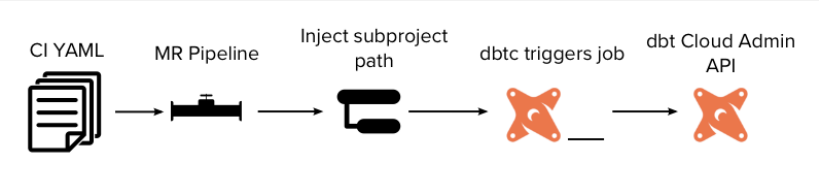


dbt-snowflake zero copy clone
upstream_prod
Redshift proxy views

Setup Slim CI in your MR Pipelines

Project federation requires a custom Slim CI solution due to conflicting databases and dbt projects in a monorepo

Make use of dbtc to connect to dbt Cloud API when running Slim CI, but configure job in CI/CD pipeline where we can set rules




```yaml
dbt_slim_ci:
  extends:
    - .matrix
  script: dbtc trigger-autoscaling-ci-job
  rules:
    - if: $CI_PIPELINE_SOURCE == "merge_request_event"
      changes:
        - $DBT_PATH/**/*
      when: manual
      allow_failure: false
```

Thank you! Questions?

Tyler Rouze

Linkedin.com/in/tylerrouze/

trouze@analytics8.com

tylerrouze.com

Building a Modern Data Team in 6 Months

Jack Blandin

VP of BI & Data Science, Fi

Linkedin.com/in/jack-blandin-19761847/

About me

- VP of Data at Fi
- Previously led Data Science & Machine Learning orgs at Wayfair and GoHealth
- Started my career as a full stack software developer at Trunk Club

Purpose of today's talk

- Share strategies for navigating real world challenges to building data teams.
- Emphasis on soft-skills.
- Walk-through of my first 6 months at Fi and how we built the data team to what it is today.
- What worked well, what didn't work well, and lessons learned.

Context

Day 1

About Fi

- The Product – Smart dog collars
  - GPS tracking
  - Activity detection
  - Walk & sleep tracking
  - Step counts
  - Rankings …
- The organization.
  - Fast-growing startup.
  - Company was ~75 people.
  - Eng was ~15 people.
  - Marketing & Finance awerere ~6 people each.
  - Data team was me and one Data Analyst.

Get the lay of the land

Weeks 1-4

Weeks 1-4

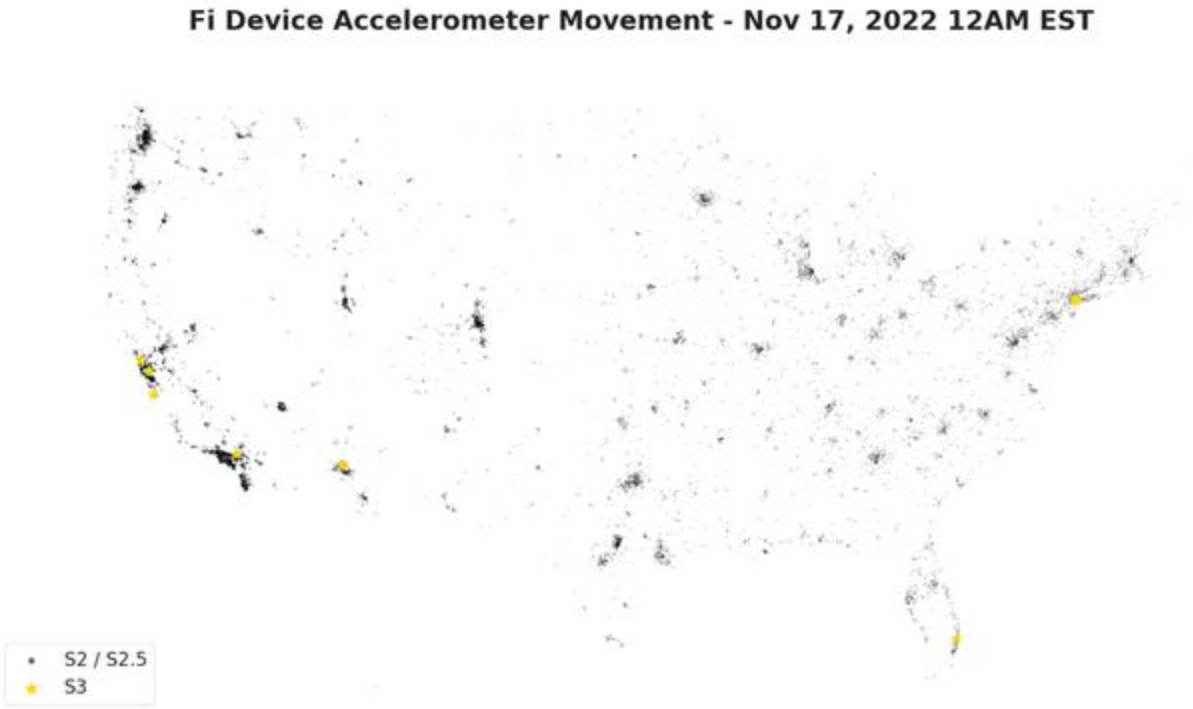
Goals

- Get the lay of the land.
- Demonstrate ability.

How?

- Understand the role of data team
- Understand the primary stakeholders and what they value
- Understand the technical process for providing solutions
- Keep an eye out for low hanging fruit
- Show off a little if there's an opportunity

Example: Fi Dogs "waking up" across the country



Demonstrate Value

Weeks 5-8

Weeks 5-8

Goals:

- Demonstrate value. Show what value the data team can provide.

How?

- Grab the low hanging fruit.
- Find a "this seems off…" situation.

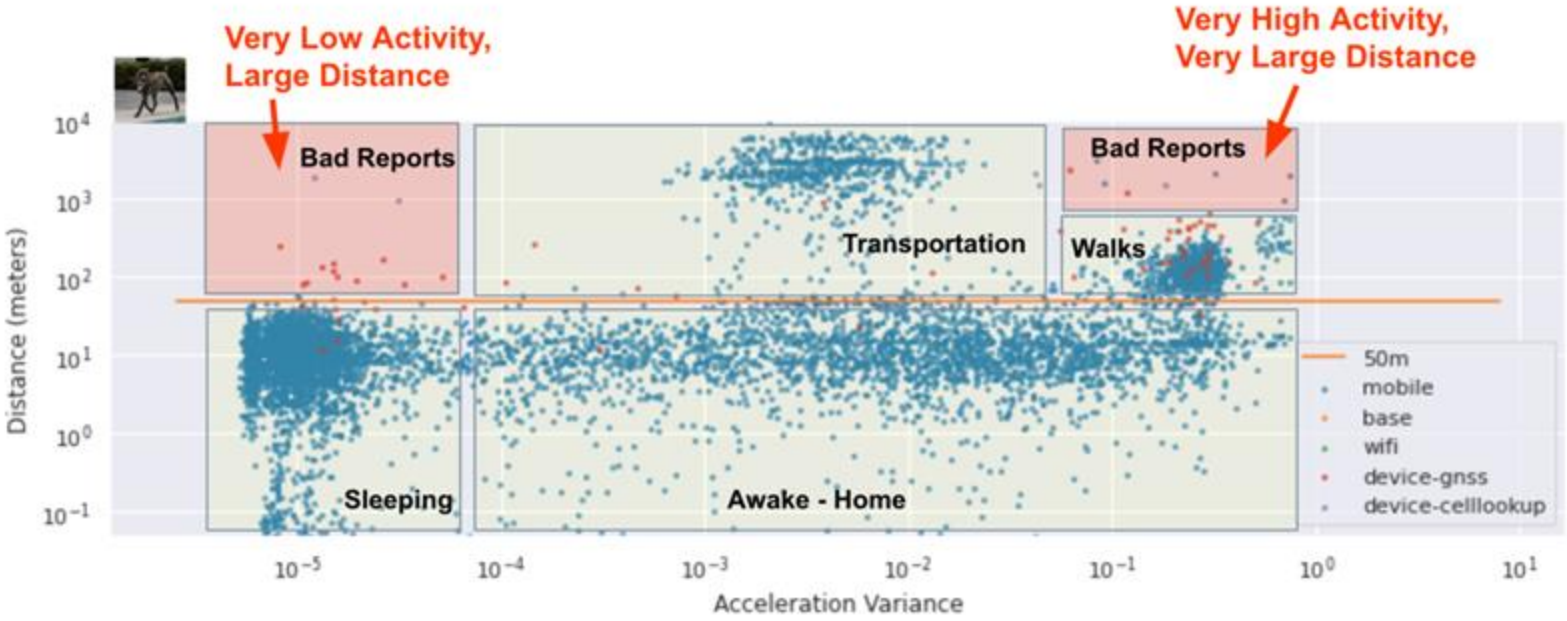
Tips

- Don't invent problems. Find something people actually care about.

Example: Identifying when GPS is incorrect

- Fun fact: GPS signals are often unreliable!
- Locating tracking is non-trivial, need to use various input signals to pinpoint exact location.
  - Remember the early days of driving navigation systems?
- How does this cause problems for a dog tracking company?
  - My phone: "Your dog has escaped!"
  - My dog: Snoring on the couch next to me.
  - Me: ☹️
- How to deal with it?
  - Talk to people!
  - Someone will have ideas for how to solve it.

Example: Identifying when GPS is incorrect



Vision & Plan

Weeks 8-10

Weeks 8-10

Goals:

- Develop vision, plan of attack, and get buy-in.

How?

- Articulate what people are needed and why.
- Articulate what tech stack is needed and why.
- Articulate what processes are needed and why.

What people are needed?

- Data Engineer.
  - Data ingestion setup and maintenance.
  - Data pipeline orchestration.
  - Real-time streaming (1 year out)
  - General SWE stuff needed by the data team.
- Analytics Engineer
  - Business savvy. Expert at the X-Y problem.
  - Can start with a business question and back into the solution.
  - Strong technical foundation, but not hyper-focused on tech stack experience.

Buy-in

Value added from Data Engineer

- Unlocks ability to analyze data coming from the hardware.
- Unlocks new features that further the margin from our competitors

Value added from Analytics Engineer

- Time to answers
- Product success tracking
- Increased ability for stakeholder teams to self-serve.

Hiring Process

Weeks 10-18

Weeks 10-18: Hiring

Goal

- Design the hiring and interview process for finding the people you need.

Analytics Engineer

- Tech challenges are embedded in business scenarios.
- Emphasize forward-thinking when problem solving (no flailing).
- Foundation over specific tools.

Tips

- Iterate!
- Resist the urge to settle
- Resist the urge to evaluate candidates on how well they'll resolve your immediate problems
- Reach the same conclusion on two separate occasions before making a final decision.

Context

A hospital emergency room is receiving an overwhelming number of patients experiencing chest pains. Chest pains are a symptom of cardiac arrest (heart attack), and therefore patients who are showing other symptoms of heart attacks should be prioritized (fast-tracked) upon entering the ER waiting room. On average, the ER is equipped to **fast-track 20%** of patients who are experiencing chest pain, allowing them to cut the waiting line. Currently, the ER's policy is to fast-track any patients who are experiencing **Type 2 chest pain (atypical angina)**. The ER is requesting that you perform analysis on patient data in order to develop a better system for fast-tracking patients so that they are fast-tracking more patients who are actually going to experience heart-attacks and therefore require critical care.

Task

You are given a dataset consisting of patient health information. This information is only for patients who came to the ER experiencing chest pains. This dataset includes 13 readings taken at the time of entering the ER, as well as the outcome (whether or not they ended up having a heart attack).

Using the provided dataset, come up a with a better decision process for fast-tracking ER patients. You may use any of the 13 inputs in the decision process.

As a first step, we suggest that you define a problem statement. What is your goal here?

What to look for

- Demonstrates forward-thinking and problem scoping.
  - "Amateurs talk strategy. Professionals talk logistics."
  - Within the first few minutes, I want the candidate to articulate their goal in a very clear manner.
- Opts for a simple approach when possible
  - I don't care that you know how XGBoost works if you choose to use it on a dataset with 303 records.
- How they collaborate.
  - Clarifying questions?
  - Do they embrace suggestions or do they get defensive?
  - Understand the intent of your questions?

Onboarding & Role Shaping

Weeks 18-24

Weeks 10-18: Onboarding & Role Shaping

Goal:

- Get the new team members familiar with the people, data, tools, and processes.

How?

- Don’t give new people new problems.
- The tasks they're solving aren't as important as the foundation they're building.

Tips

- Shape the roles to the people and to the team, not the other way around.
- Be clear about expectations, even while roles are being shaped.

Hit the gas

Weeks 24-26

Weeks 24-26: Hit the gas

Goals

- Deliver value
- Increase the team's capacity to deliver value

How?

- Infrastructure in lock-step with value-add
- Delegate
- Encourage collaboration

Tips

- "Vacuum"
- (Soft) deadlines
- Celebrate wins!

Closing Remarks

Closing Remarks

The Fi data team today

- The team:
  - Myself (Chicago)
  - Staff Data Eng (Maryland)
  - Staff Analytics Eng (Los Angeles)
  - Data Analyst (NYC)
  - +2 head count for Data Analysts (NYC)
- The stack:
  - AWS
  - Databricks (delta lake)
  - PySpark
  - dbt
  - Tableau (for now)

Closing Remarks

Final Tips

- Consistently provide value. Don't try and flip the iceberg.
- Leverage people's expertise when doing data analysis.
- Bring experience, but don't bring baggage.
- If there's one thing to make sure you do well – focusing on hiring will give you the biggest return on investment.

Thank you! Questions?

Jack Blandin

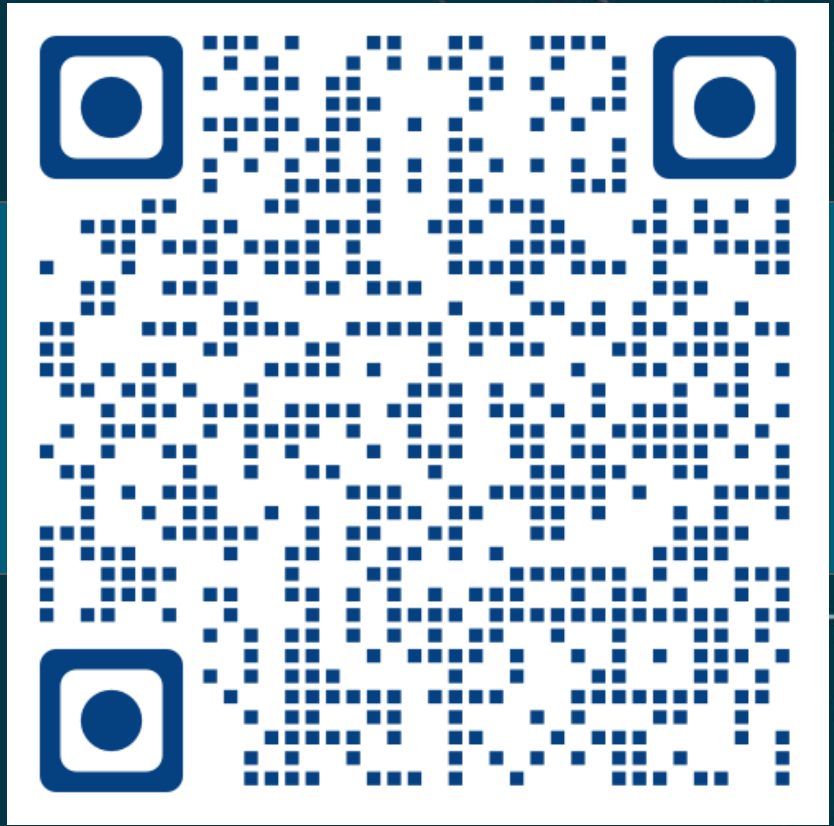
Linkedin.com/in/jack-blandin-19761847/

Thank you!

Stay tuned for upcoming events and other content

Chicago dbt Meetup Group
#local-chicago Slack Channel

dbt

Before you leave...

Please share your feedback with us!