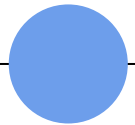
goto;

GOTO
CHICAGO 2023

#GOTOchgo

Bridging the gap: How data and software engineering teams can work together to ensure smooth data integrations



Sam Bail @spbail

GOTO Conf Chicago, 2023

The TL;DR

- Integrating data from a new source into your data pipeline isn't just "plug n play"
- I'll cover some of the most important points to discuss when software engineering and data teams work together
- This talk is for data people and software engineers that work with data teams

About me

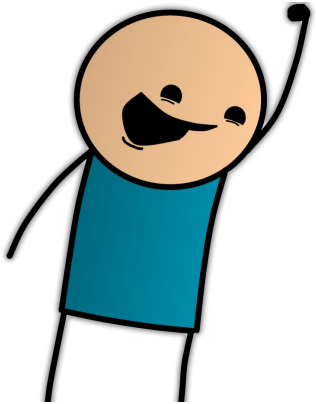
- I’m Sam Bail (she/her) @spbail
- German based in NYC
- I’m a one-stop data shop
- Currently: Data Engineer at Collectors
- Fun fact: I run a non-alcoholic pop-up bar @thirdplacebarnyc

Outline

- The problem
- Areas to cover:
  1. Logistics
  2. Infrastructure
  3. Data model
  4. Application and data flow
  5. Data contracts
- Wrapping up

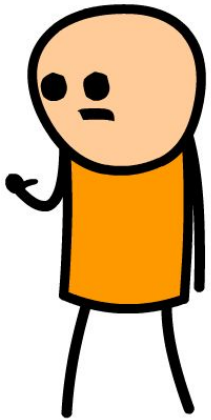
The problem

"We're launching this awesome new feature next month! And we need analytics from day 1! Let's GOOO!"



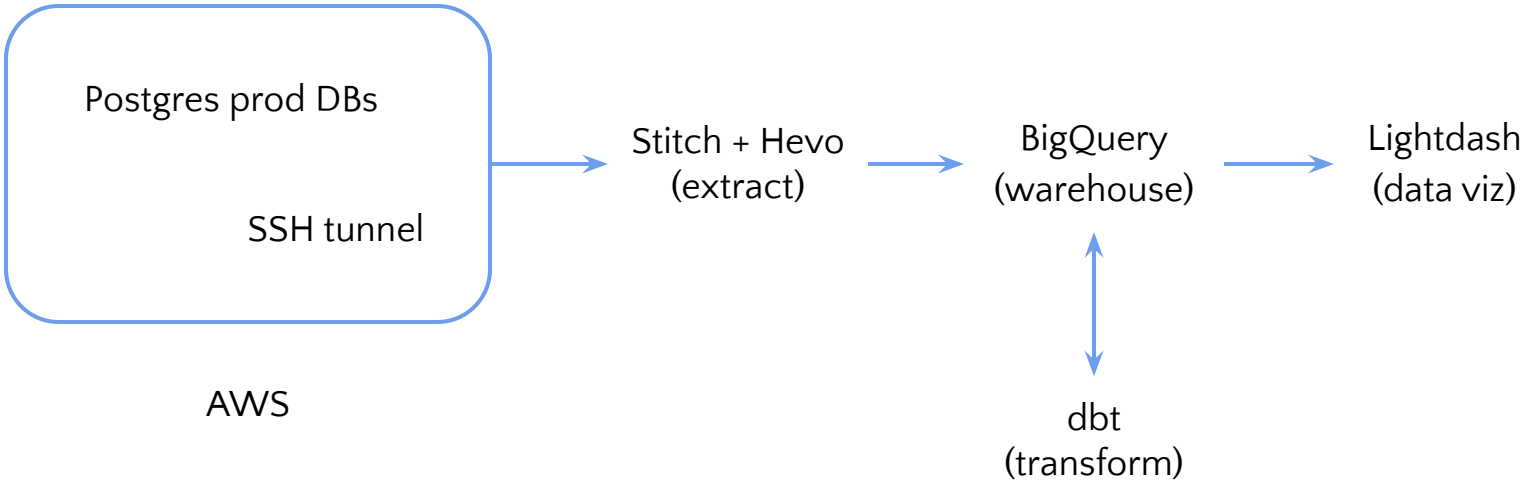
Product manager

HOLD ON! Lemme talk to the software engineering team first and see what their data architecture looks like...



Data team

Our (modern) data stack



1. Logistics

Who does what where when?

Keep a running notes doc.

1. Logistics

Who does what? Engineering lead, product lead? Establish a connection.

1. Logistics

How do we communicate?

Standing meetings? Slack? Email? Docs? Jira?

1. Logistics

What do we want to measure & when?

On day 1, week 1, later? Is the data actually available?

1. Logistics

How do we keep the software engineering team in the loop on analytics? (Show, don't tell!)

1. Logistics

2. Infrastructure

How do we do the plumbing?

Where is the data hosted? What type of data storage is it? Can our ETL tool actually handle this? Do we need an SSH tunnel?

2. Infrastructure

Are there dev and prod instances? Will we access prod or read replicas? Do we need any kind of write access (e.g. for temp tables, views)?

2. Infrastructure

Do we need a read-only account? Service accounts? Personal accounts? How will credentials be shared?

2. Infrastructure

When will we get access to the data?

To dev? Prod?

2. Infrastructure

3. Data model

What to expect when you're expecting (data)

What does the data schema look like? Is there data documentation? Who owns maintaining it and communicating changes?

3. Data model

Will data constraints be enforced in any way? E.g. foreign key relationships, NULL values, default values, JSON schemas...

How do we handle timezones? (LOL) Currency?

3. Data model

Are we actually storing everything we want to measure?

3. Data model

4. Application and data flow

What happens when I click here?

How and when are records created and fields populated?

4. Application and data flow

What actions cause records to be modified? How are modification events logged? Do we have “updated at” fields or log tables?

4. Application and data flow

How are deletions being handled? Do we have hard deletes or soft deletes? Will “old” data be archived after some time?

4. Application and data flow

Is data being migrated from a legacy application? Will there be anything different from new data that's being created?

4. Application and data flow

Will there be (realistic) test data we can develop against? Will there be test data left in the production system?

4. Application and data flow

5. Data contracts

How do we keep this working?

How do we document all these things we just agreed on?

5. *Data contracts*

And how do we enforce them going forward without requiring too much human input?

5. *Data contracts*

Will there be database constraints or any kind of testing as part of CI/CD on the data producer side (not just the data consumers)?

5. *Data contracts*

How do we communicate changes to the data?
Who needs to be informed?

5. *Data contracts*

What's the procedure in case something breaks? How do we report, and what's the SLA for fixes?

5. *Data contracts*

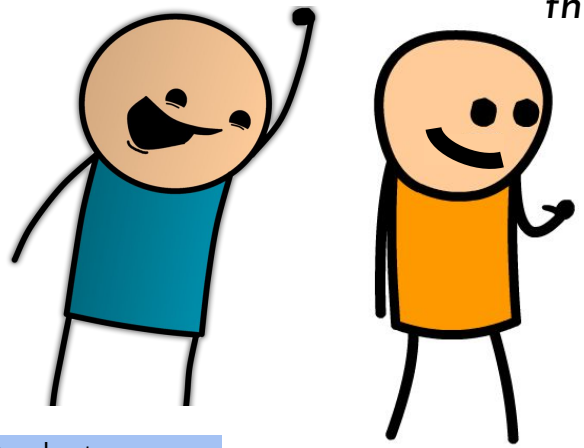
Wrapping up

- Integrating data from a new source into your data warehouse isn't just "plug n play"
- There is an infinite number of questions to consider. You will probably miss something.
- The key is connection and context between teams.

The end

Look at this awesome new feature! And the dashboard to track all these cool metrics!

Well it's not everything you asked and it was a bit bumpy getting there, but it works! Go team!



Product manager

Data team

Resources

- Feel free to reach out to @spbail on Twitter
- Or email sbail@collectors.com
- This talk is based on a blog post: https://sambail.com/2022/09/19/analytics-from-day-1/

goto;

Don’t forget to
vote for this session
in the **GOTO Guide app**