



&



better together

Sam Debruyn



Data Platform Next Step 2023

dataroots

Who is Sam?

👛 tech lead data & cloud at **dataroots**

📅 4 years in data

📅 9 years in software / architecture / cloud

❤️ FOSS, modern data stack, Azure

👤 maintainer of dbt adapters for all MS SQL products & top contributor to dbt-core

🎸 concerts, festivals, guitar

dataroots



**Have you ever
asked yourself...** 

**When was this table
last refreshed? **

**Why is this view
here? What does it
do? 🙄**

**Are the values in
this column
expected?**

**Where is the code
for this view? 🔍**

**Why did we change
this view/query?** 🤔

**Who changed this
view/query? 🤔**

**Wait, how many
levels deep did this
subquery start? 🤔**

**How long is this SQL
query? I need an 
aspirin!**

Where is Dave?



**He needs to log on to
SSMS on prod to
deploy the new
code!** 

**Dave is on
holiday???** 🌴 🍹

**What should we do
now? 🥲**



What we don't want

- Complex/obscure Python/Scala/DAX logic
- UI building blocks without clear code-behind
- Unexpected behavior in the code
- Unrealistic data
- Answering the same questions about a table all the time
- Analytics is only done by small team inside the company



What we want

- SQL with superpowers
- Clear changes and diffs in git
- Automated testing of code
- Automated data quality monitoring
- Automated deployments
- Great documentation of the data model
- Everyone can become an analytics engineer

dbt?

Dialectical Behavior Therapy

~~Dialectical Behavior~~ Therapy



"Analytics Engineering"

Applying software engineering best practices on data modeling and warehouse/BI development

Behind The Hype - Is Analytics Engineer a Real Job



Advancing Analytics

22.2K subscribers

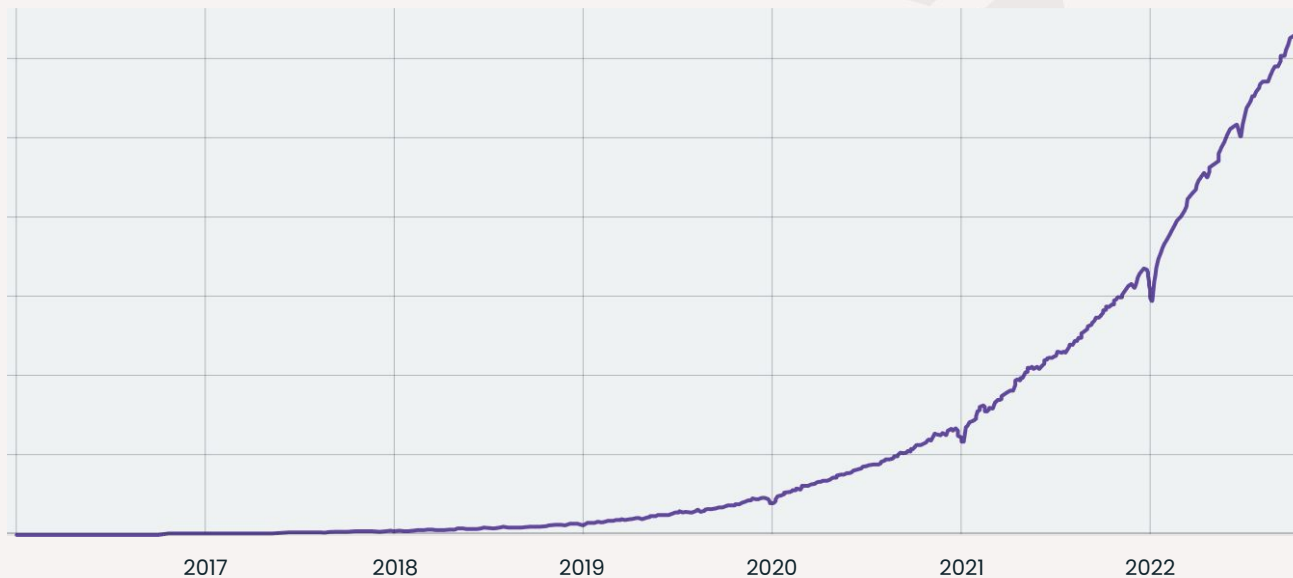


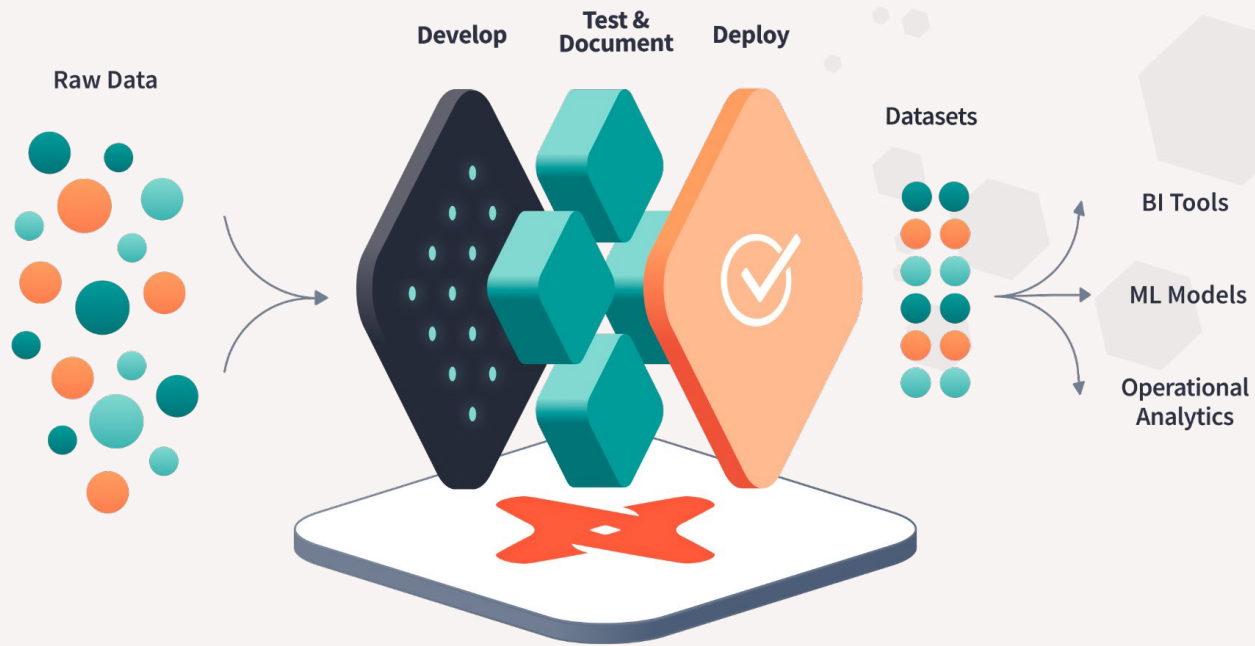
Subscribed

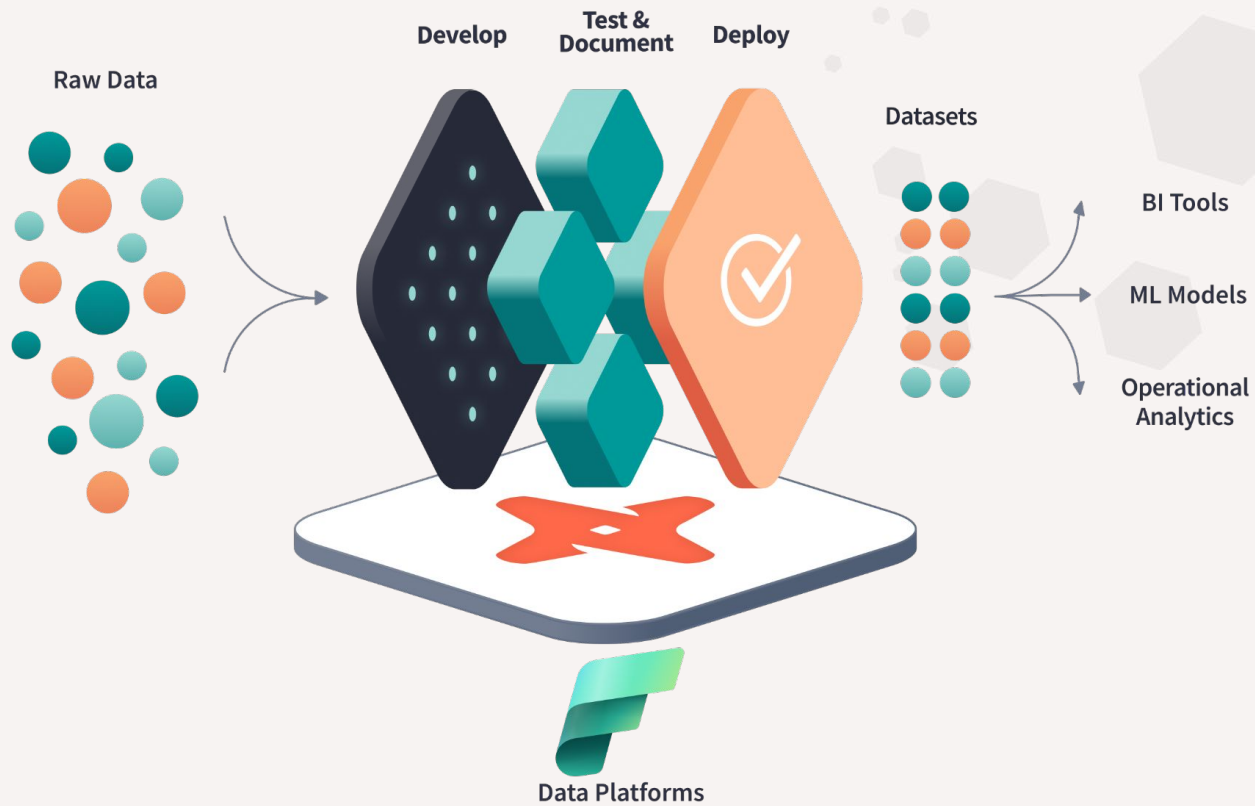


dataroots

Q3 2022: 16000 companies using dbt







Data Source



Shortcut Enabled



Structured /
Unstructured

Ingestion



Mounts



Pipelines &
Dataflows

Store



Data Warehouse



Transform

Expose



PBI



Warehouse

dataroots

What dbt is not

Data warehouse/lakehouse

Compute engine

Alternative to a Spark cluster (can replace your Spark code though)

Schema migration tool

Drag & drop tool

Compatibility

Adapters as bridges between
dbt and the data warehouse



Starburst



DuckDB





dbt Core















- Open source & free to use
- Complete dbt experience with all capabilities in regards to transformations, database engine compatibility



dbt Cloud

- Free tier: 1 user
- Teams: starts at 100 USD per user
- Everything in dbt Core for the most common engines
- Browser-based IDE
- CI/CD & job scheduling
- Hosted data docs
- Easy setup & configuration
- ...

dbt support

Engine	Supported by	dbt Core	dbt Cloud
SQL Server 2017 or newer	 Community		
Azure SQL (DB/MI)	 Community		
Azure Synapse Dedicated SQL Pools	 Community   Microsoft		
Fabric Data Warehouse	 Microsoft		



How does it work?

dataroots

Demo 🙌

packages.yml

Package hub



```
packages:  
  - package: google/fhir_dbt_analytics  
    version: 1.0.1  
  - package: fivetran/linkedin  
    version: 0.7.0
```

tests/assert_payment_amount_positive.sql



```
select
    order_id,
    sum(amount) as total_amount
from {{ ref('fct_payments') }}
group by order_id
having not(total_amount >= 0)
```


models/src_orders.yml




```
version: 2
sources:
  - name: orders
    database: shop
    freshness:
      warn_after: {count: 12, period: hour}
      error_after: {count: 1, period: day}
    loaded_at_field: _elt_loaded_at
```

Accomplish great things




Version controlled and reproducible

 Collaboration within the team & other teams




Built-in docs & lineage

 Know and understand your data




Test code & data

 Deploy & run with confidence



Modular & easy to use

 Easy to extend and maintain

dataroots



There is more

- Implement SCD with *snapshots*
- *Incremental* loads
- Parametrize code with *Variables*
- *Hooks & operations*
- Run *Python models* through Spark
- Manage access with *grants*
- Track dataset usage in BI & ML with *exposures*
- ...


Where to go next?


 <https://www.getdbt.com/dbt-learn/>

 `pip install dbt-fabric==1.4.0rc1`

 community:

 50k+ Slack members

 lots of meetups all around the world

 @s_debruyne /  sam@dataroots.io

dataroots