Sawasdee dbt

The Data Transformation Tool that Makes Analytics Engineering Easier

Kan Ouivirach

<DEV>
MOUNTAIN
TECH FESTIVAL

Kan Ouivirach

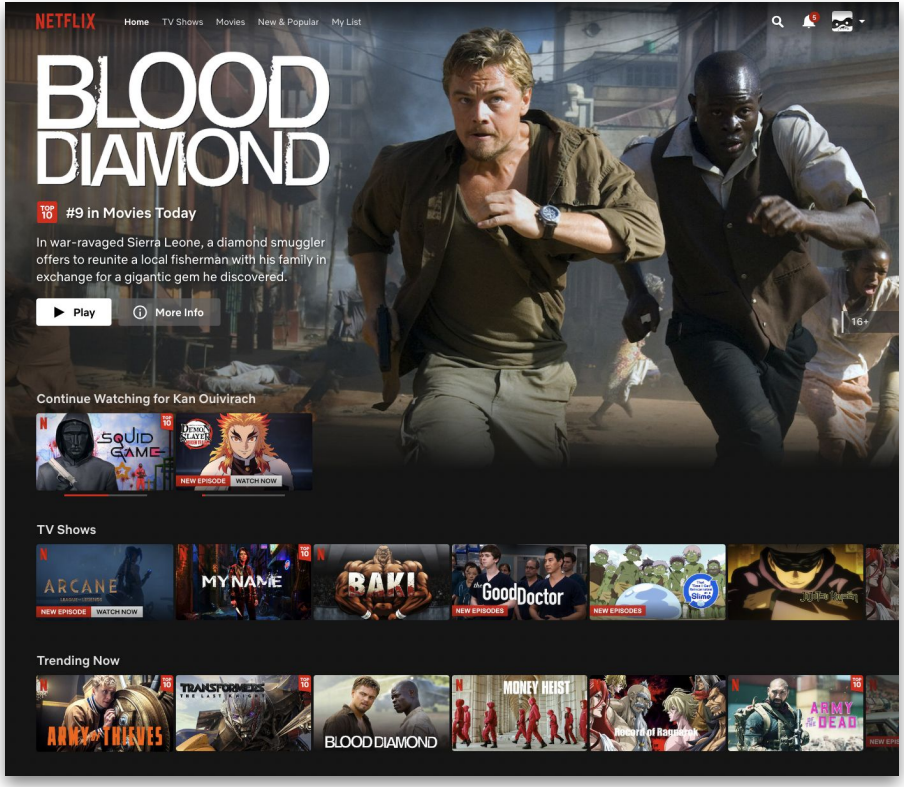
- Data Architect & Engineer
- Astronomer Apache Airflow Certified Expert
- Contributor in Apache Airflow
- Community Organizer
  - Data Engineer Cafe
  - Data Council Bangkok
  - PyCon Thailand

Outline

- What is Analytics Engineering (AE)?
- Why does **dbt** make AE easier and better?
- How can we use it?

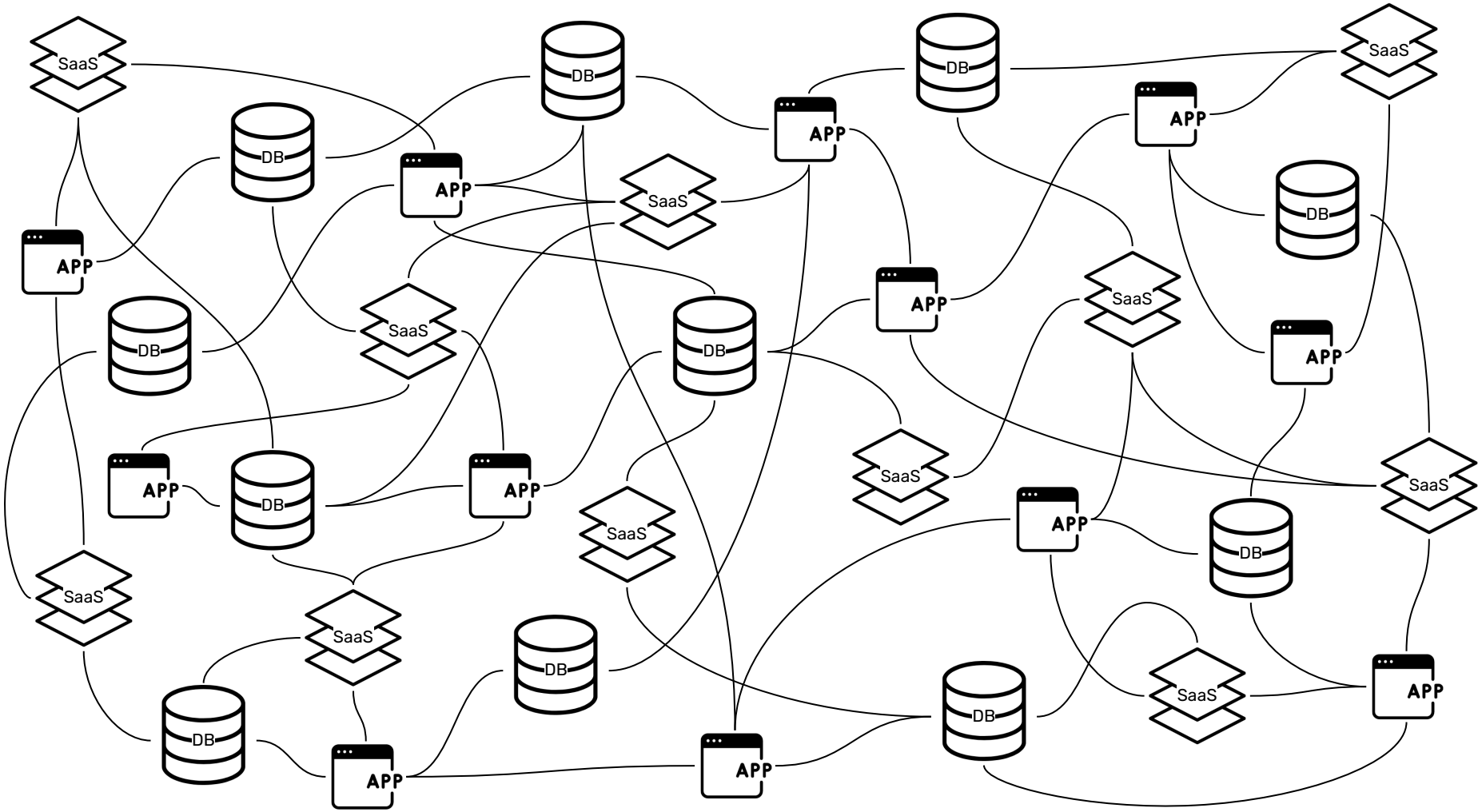
Companies are **becoming** software

Photo by Peter Nguyen on Unsplash



Rich frontend customer experiences

Complex backend operations


SaaS
DB
DB
SaaS
DB
APP
APP
APP
SaaS
DB
APP
APP
SaaS
DB
SaaS
DB
APP
APP
DB
APP
SaaS
SaaS
SaaS
SaaS
APP
DB
SaaS
SaaS
DB
DB
SaaS
APP
DB
APP
APP

Companies are **becoming** software

How do we build a “sustainable” company in software?

Photo by Peter Nguyen on Unsplash

Software is essentially complex..

So is analytics.. 🥲

Production software requires software engineering

So does analytics.. 😤

Have you experiencing these in data analytics?

- Write the same SQL query for the 10th time..
- Saved lots of personal queries in a data warehouse
- Never reuse
- Write once, forget forever..
- No collaboration in the team
- Slowly start to come back to the old queries
- Broken dashboard and take time to find issues with data
- No idea when data is stale..
- No idea where data come from..
- Difficult to scale the analytics
- etc.

We should fix!

What is Analytics Engineering?

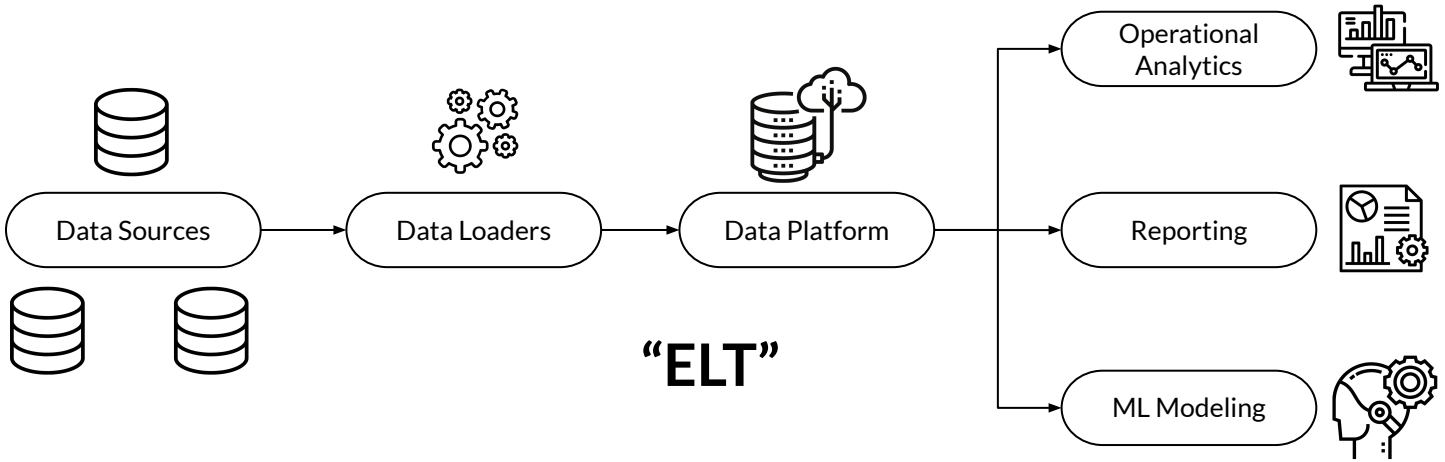
Analytics Engineering

Applying **software engineering** best practices to the analytics code used to build and feed the data sets

- Transforming
- Testing
- Deploying
- Documenting
- Version Controlling
- Continuous Integration
- etc.

Why does it become more popular?

Modern Data Stack



- Aim to lower the technical barrier to entry for data integration
- A suite of tools used for data integration
- Hosted in the cloud
- Built with analysts and business users in mind

Traditional Data Team

Data Engineers

- Build and maintain data platform & infrastructure
- Build and manage data pipeline orchestration
- Optimize data warehouse performance
- Deploy machine learning models
- Build data tools

Data Analysts

- Build dashboards & reports
- Work with business users to understand data requirements
- Find insights in data

Modern Data Team

Data Engineers

- Build and maintain data platform & infrastructure
- Build and manage data pipeline orchestration
- Optimize data warehouse performance
- Deploy machine learning models
- Build data tools

Analytics Engineers

- Own the transformation of raw data up to the BI layer
- Provide clean, transformed data ready for analysis to business users
- Model data in a way that empowers business users to answer their questions
- Apply software engineering practices to analytics
- Deliver well-defined, transformed, tested, documented, and code-reviewed data sets

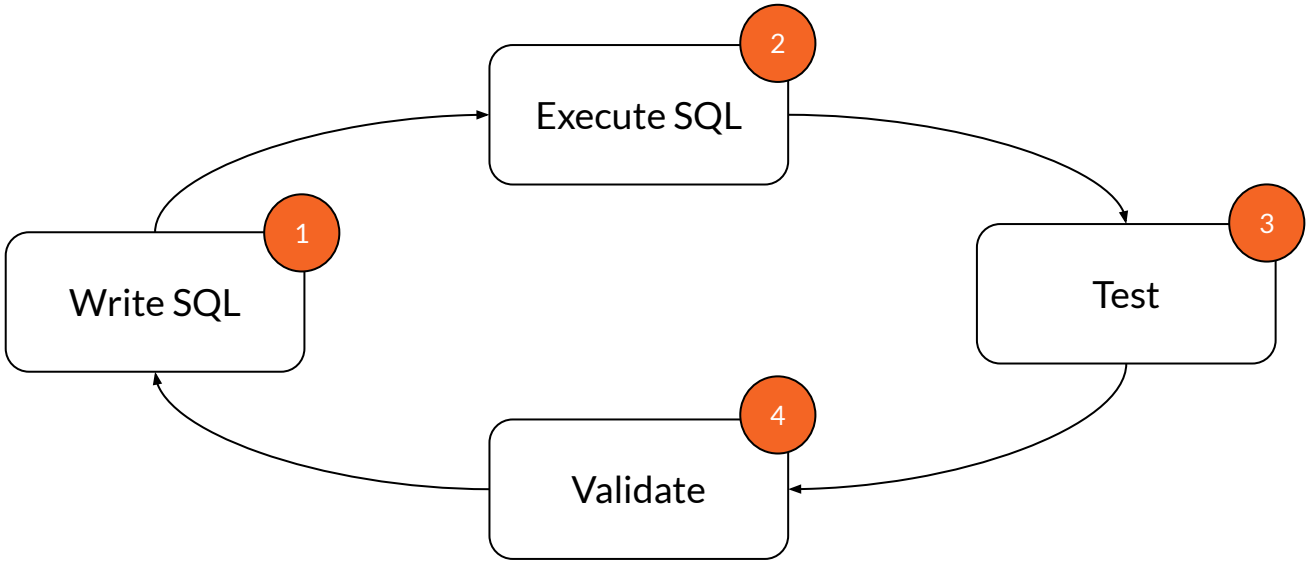
Data Analysts

- Build dashboards & reports
- Work with business users to understand data requirements
- Find insights in data

Analytics Engineers Care About

- How to provide clean and transformed data to answer an entire set of business questions?
- What is the good naming convention for tables in the data warehouse?
- How to notify when a problem in the data found before a business user finds a broken dashboard?
- What do analysts or other business users need to understand about this table to be able to quickly use it?
- How can I improve the quality of data as its produced, rather than cleaning it downstream?

Analytics Workflow



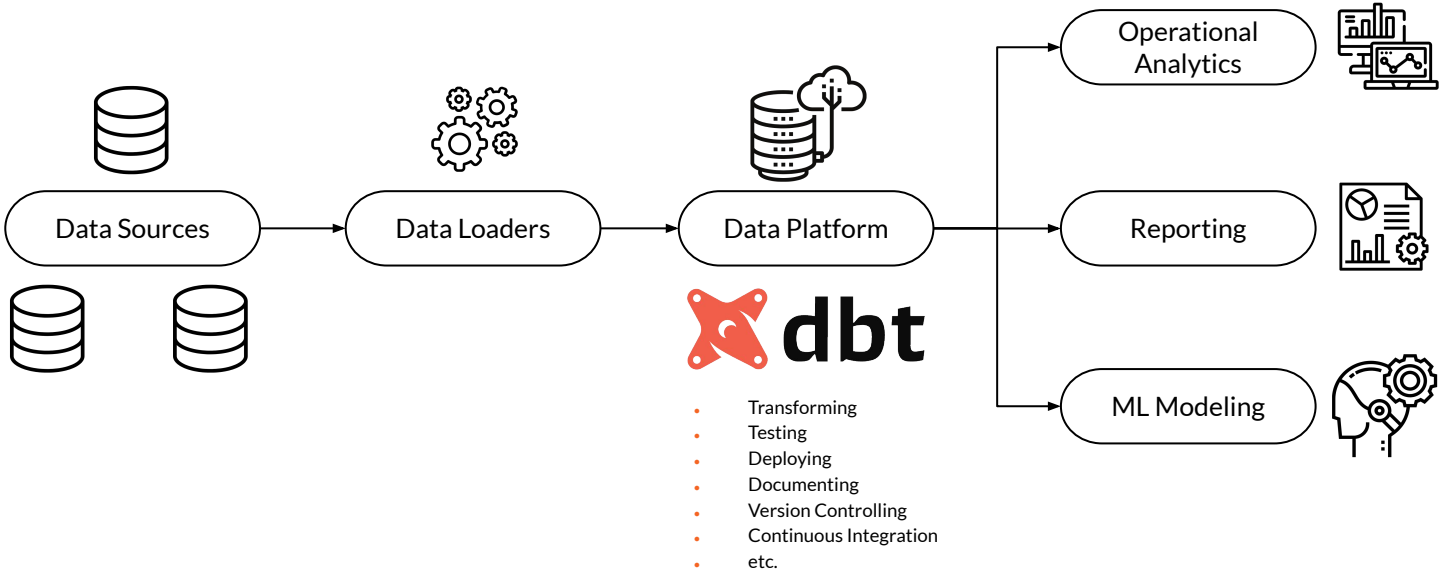
The “T” in ELT

Why does **dbt** make AE easier and better?

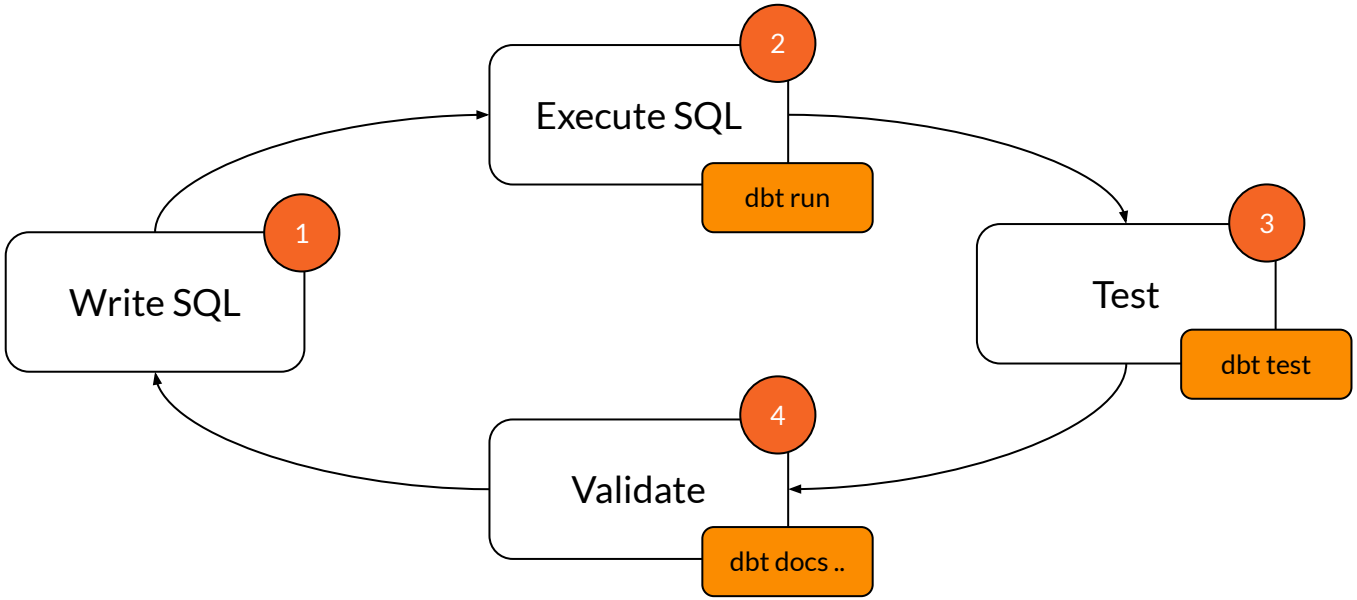
What is dbt?

- dbt (**d**ata **b**uild **t**ool) developed by dbt Labs (formerly the Fishtown Analytics)
- Open source, CLI tool with SQL based solution for the “T” in ELT
- Take care of dependencies, compilation, and materialization in run time
- Empower data teams to leverage software engineering principles to analytics workflow

Modern Data Stack



Analytics Workflow with dbt



Why dbt?

- SQL on Steroids
- Configuration as Code
- Automated Testing
- Environments & Automation
- Auto-Generated Docs & Lineage

SQL on Steroids

- Combine SQL with Jinja
  - Macros to apply the post-hook PII data masking
  - Iterate over multiple objects
  - Pretty much anything..
  - etc.
- Control structures & environment variables
- DRY principle
- Think functions, not stored procedure

dbt Models

- Models are just `select` statements that transform data
- File `.sql` in `models` directory

```sql
SELECT
  sp.int AS payment_id,
  sp.orderid AS order_id,
  jso.user_id AS user_id,
  sp.paymentmethod AS payment_method,
  sp.status AS payment_status,
  sp.amount
FROM
  stripe_payment sp
JOIN
  jaffle_shop_order jso
ON
  sp.orderid = jso.id
```

Configuration as Code

- Create tables, views or custom objects
- Choose how to partition & cluster objects
- Configure at an object or dataset level

Automated Testing

- Transformations only useful if correct
- Help define automated tests
- Build confidence in the insights and when making changes

About Testing in dbt

- Two types
  - Generic (`unique`, `not_null`, `relationship`, `accepted_values`)
  - Singular
- Tests are assertions about the models
- When assertion is true, the test passes
- Tests are just select statements!

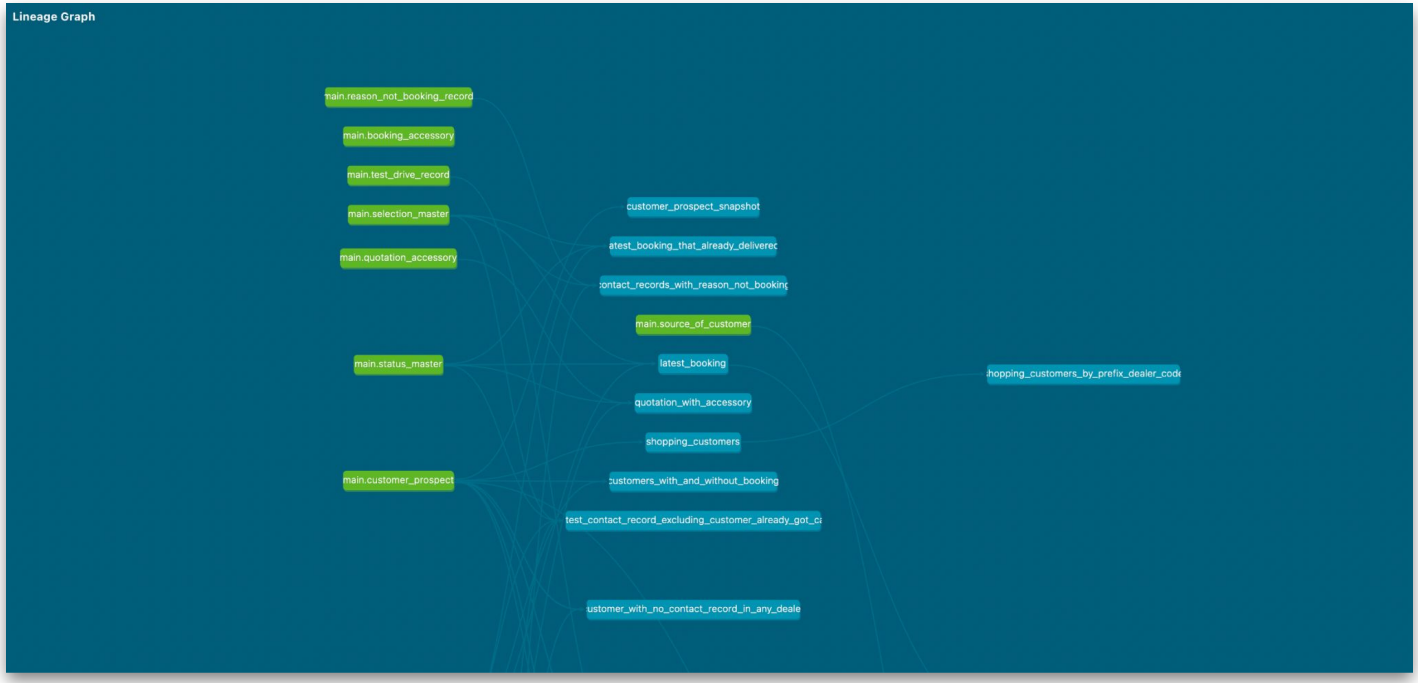
Environments & Automation

- Help manage different environments
- Use the same code & run in any environment
- Automate using tool of choice, e.g., Airflow or Jenkins
- Help mitigate against config drift

Auto-Generated Docs & Lineage

- Built-in documentation/website
- Visualize object relationships — Lineage
- Serve as a data catalog
- Can be hosted anywhere

Real Example of Lineage Graph



dbt in Short

- Remove the overhead of data transformation
- Encourage software engineering principles in data
- Less effort, more insights & business value
- Help build a robust data warehouse
- Promote collaboration

How we can use it?

What We Need to Use dbt

1. Data warehouse or database
2. dbt project
3. Connection to the data warehouse or database
    a. dbt profile
4. Command

How dbt Works

1. Connect to the data warehouse (via a profile/connection)
2. Parse the dbt project
3. Wrap the models in the appropriate DDL/DML (e.g., `create table as`)
4. Execute the code to build models in target schema

Demo

https://github.com/zkan/hello-dbt/

Conclusion

- Analytics does requires software engineering practices
- Your analytics engineering workflow can be easier and better with dbt
- dbt helps build a scalable process of developing code with built-in testing of code, implicit DAG, online data catalog and lineage

“That’s the future of analytics”

<DEV>
MOUNTAIN
TECH FESTIVAL

Data Engineer Cafe

https://discuss.dataengineercafe.io