What is dbt?

INTRODUCTION TO DBT

Mike Metzger
Data Engineer

dbt defined

- "data build tool"
- Primarily handles the T in ELT (sometimes ETL)
- Allows easy switching between data warehouses
- Ideal for teams with different needs
- Provides source / code control

What does dbt do?

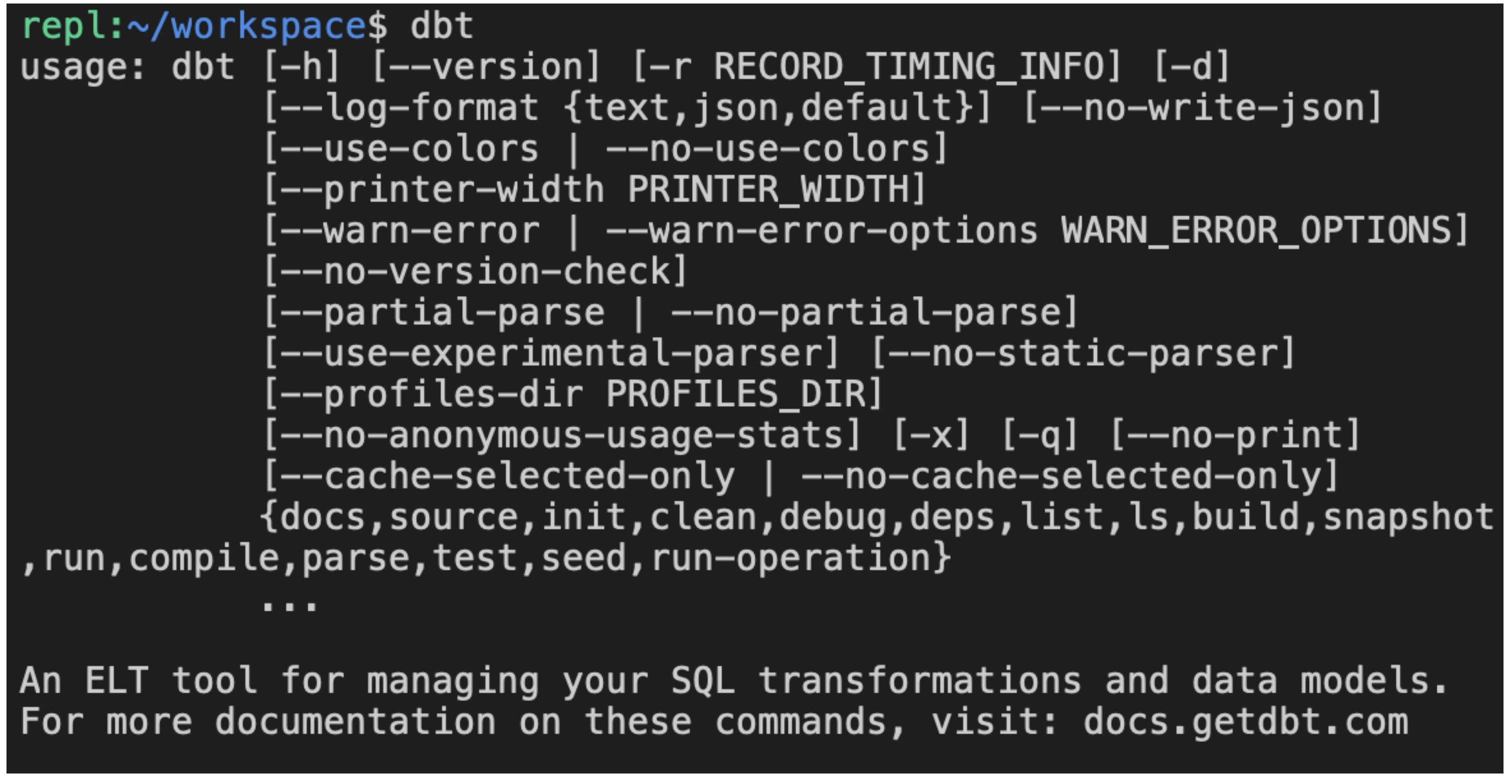
- Primarily defines data models and transformations using SQL
- Newer versions can use Python
- Translates between SQL dialects
- Can define relationships between data models
- Runs the data transformation process(es) as requested
- Can test for data quality requirements

What does dbt look like?

- Command-line tool `dbt`
- Also known as `dbt-core` , open-source
- Adapters provide connections to different data warehouses
  - dbt-snowflake
  - dbt-bigquery
  - dbt-sqlserver
- dbt Cloud

dbt subcommands

- `dbt` has several subcommands
  - `dbt` or `dbt -h` - Shows help content
    - `dbt <subcommand> -h` - help for subcommand



```
repl:~/workspace$ dbt
usage: dbt [-h] [--version] [-r RECORD_TIMING_INFO] [-d]
           [--log-format {text,json,default}] [--no-write-json]
           [--use-colors | --no-use-colors]
           [--printer-width PRINTER_WIDTH]
           [--warn-error | --warn-error-options WARN_ERROR_OPTIONS]
           [--no-version-check]
           [--partial-parse | --no-partial-parse]
           [--use-experimental-parser] [--no-static-parser]
           [--profiles-dir PROFILES_DIR]
           [--no-anonymous-usage-stats] [-x] [-q] [--no-print]
           [--cache-selected-only | --no-cache-selected-only]
           {docs,source,init,clean,debug,deps,list,ls,build,snapshot
,run,compile,parse,test,seed,run-operation}
           ...

An ELT tool for managing your SQL transformations and data models.
For more documentation on these commands, visit: docs.getdbt.com
```

dbt subcommands - part 2

- `dbt init` - Creates new dbt projects
    - `dbt run` - Runs the data generation / transformations

dbt subcommands - part 3

- `dbt test` - Used to test data quality
  - `dbt debug` - Can check connections to data warehouses
  - Many others

Who is dbt for?

dbt is designed for any users that need to transform data:

- Data Engineers
- Analytics Engineers
- Data Analysts

Rarely:

- Data scientists
- ML Engineers
- Sales / Finance / C-Level

[1] Photo by Christina @ wocintechchat.com on Unsplash

Let's practice!

INTRODUCTION TO DBT

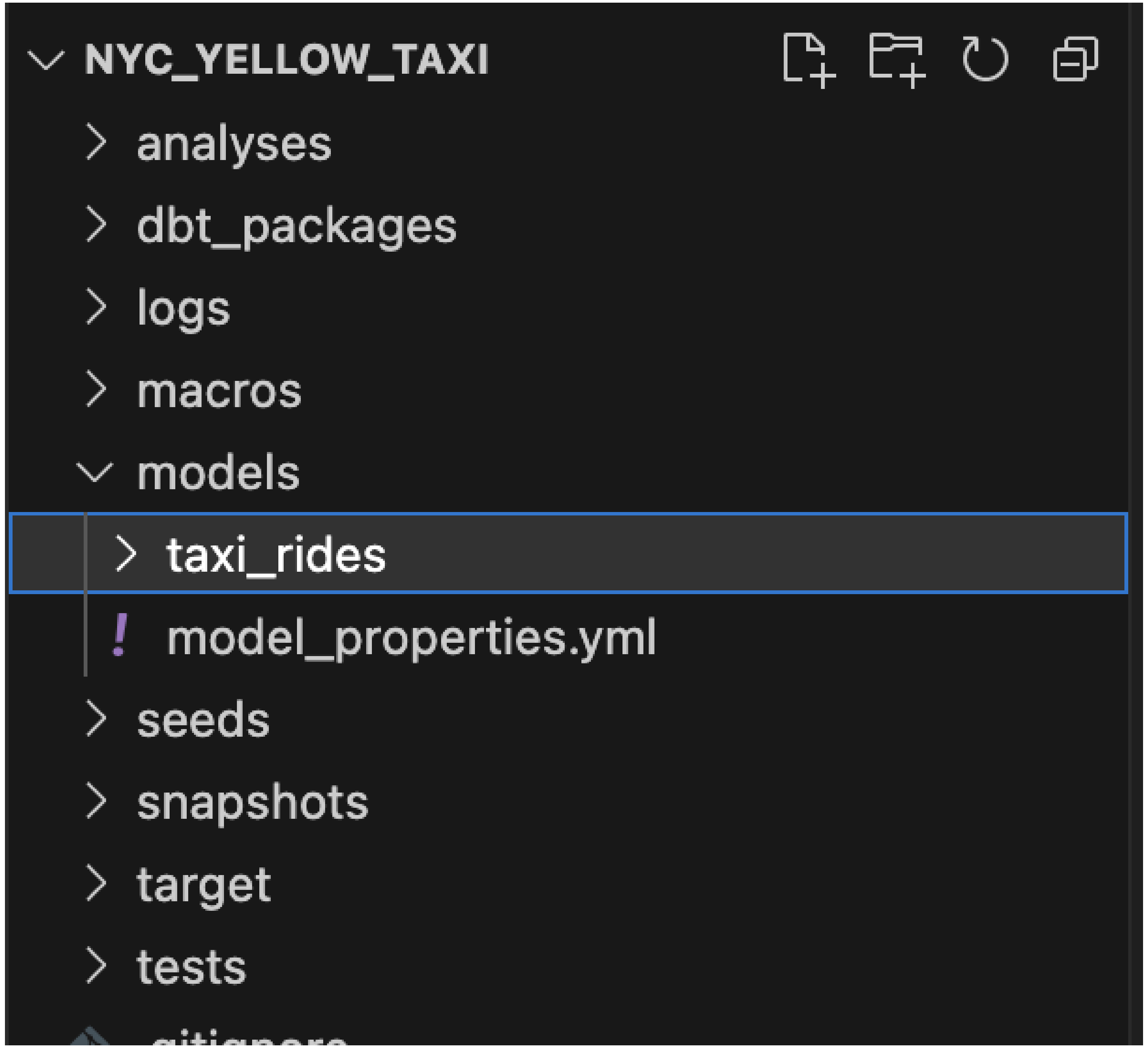
Creating a dbt project

INTRODUCTION TO DBT

Mike Metzger
Data Engineer

What is a dbt project

- Encompass components for working with data in dbt
  - Project configuration
  - Data sources & destinations
  - SQL queries
  - Templates
  - Documentation
- Implemented as a folder structure



```
v NYC_YELLOW_TAXI
  > analyses
  > dbt_packages
  > logs
  > macros
  v models
    > taxi_rides
    ! model_properties.yml
  > seeds
  > snapshots
  > target
  > tests
  .gitignore
```

How to create a new project

- Use the `dbt init` command
  - Asks the name of the project
  - Asks which database / data warehouse type
- Can consolidate with `dbt init <projectname>`
- Creates the top level project folder and all needed structure

```
repl:~/workspace$ dbt init
16:38:37  Running with dbt=1.4.1
Enter a name for your project (letters, digits, underscore): test_project
Which database would you like to use? [1] duckdb
Enter a number: 1
...
```

Defining configuration with project profiles

- A profile represents a given deployment scenario
  - Development
  - Staging / Test
  - Production
- A dbt project can have multiple profiles
- Defined in the `profiles.yml` file

```
marketing_campaign_results:
  outputs:
    dev:
      type: duckdb
      path: dbt.duckdb
    prod:
      type: snowflake
      ...
  target: dev
```

YAML

- Yet Another Markup Language
- Text file, but spacing matters (like Python)
- Used in many development scenarios for configuration
- Rules can be tricky, mainly keep entries lined up as in examples

```
marketing_campaign_results:
  outputs:
    dev:
      type: duckdb
      path: dbt.duckdb
    prod:
      type: snowflake
      ...
  target: dev
```

DuckDB

- Open-source serverless database
    - Similar to sqlite
- Designed for analytics
- Vectorized (meaning *FAST*)
- Easy to use
- dbt-duckdb

Let's Practice!

INTRODUCTION TO DBT

Working with a first project

INTRODUCTION TO DBT

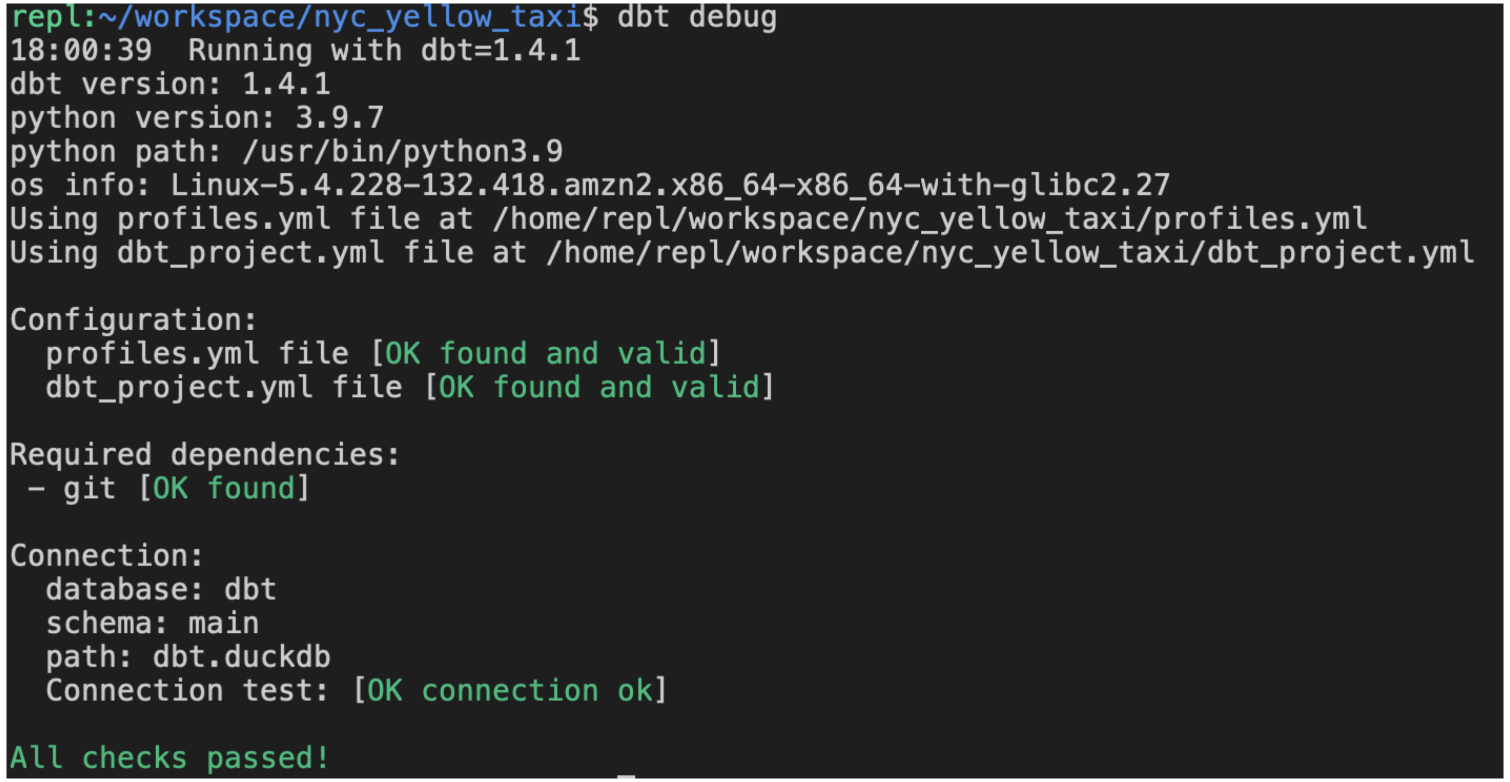
Mike Metzger
Data Engineer

Workflow for dbt

1. Create project (`dbt init`)
2. Define configuration (`profiles.yml`)
3. Create / use models / templates
4. Instantiate models (`dbt run`)
5. Verify / Test / Troubleshoot
6. Repeat as needed

Verifying database connections

- `dbt debug`
- Verifies connectivity to databases / data warehouses
- Data warehouse must be created / accessible first
- If non-existent, try `dbt run` first



```
repl:~/workspace/nyc_yellow_taxi$ dbt debug
18:00:39  Running with dbt=1.4.1
dbt version: 1.4.1
python version: 3.9.7
python path: /usr/bin/python3.9
os info: Linux-5.4.228-132.418.amzn2.x86_64-x86_64-with-glibc2.27
Using profiles.yml file at /home/repl/workspace/nyc_yellow_taxi/profiles.yml
Using dbt_project.yml file at /home/repl/workspace/nyc_yellow_taxi/dbt_project.yml

Configuration:
  profiles.yml file [OK found and valid]
  dbt_project.yml file [OK found and valid]

Required dependencies:
 - git [OK found]

Connection:
  database: dbt
  schema: main
  path: dbt.duckdb
  Connection test: [OK connection ok]

All checks passed!
```

dbt run

- Run whenever there model changes
- Or when the data process needs to be materialized
- Output provides many details on the success or failure of the various steps
- Materialized == Transformations into tables / views

Table vs View

Tables:

- Objects within a database / warehouse that hold data
- Take up space within the database
- Content only updated when changed

Views:

- Queryable like a table, but hold no information
- Are usually defined as a select query against another table or tables
- Content generated with each query

Let's practice!

INTRODUCTION TO DBT