Welcome

Taipei dbt Meetup

Host: GDG Taipei + dbt Taipei

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

dbt

Join dbt Community Slack

getdbt.com/community

add #local-taipei

Kevin Chien

MSD

Data Analyst

"A data analyst working on developing metics and solving business problems with stakeholders. Passionate about data-driven and data modeling"

✉ h.h.chien@yandex.com

in linkedin.com/in/hhchien/

Agenda

What is dbt ?

- What is transformation
- What is dbt
- What does dbt provide

Why people love dbt

- Why organizations embrace dbt
- Why Data engineer love dbt
- Why I love dbt

How to get started ?

- Prerequisites
- Installation and guide
- Become a pro

Agenda

What is dbt ?

- What is transformation
- What is dbt
- What does dbt provide

Why people love dbt

- Why organizations embrace dbt
- Why Data engineer love dbt
- Why I love dbt

How to get started ?

- Prerequisites
- Installation and guide
- Become a pro

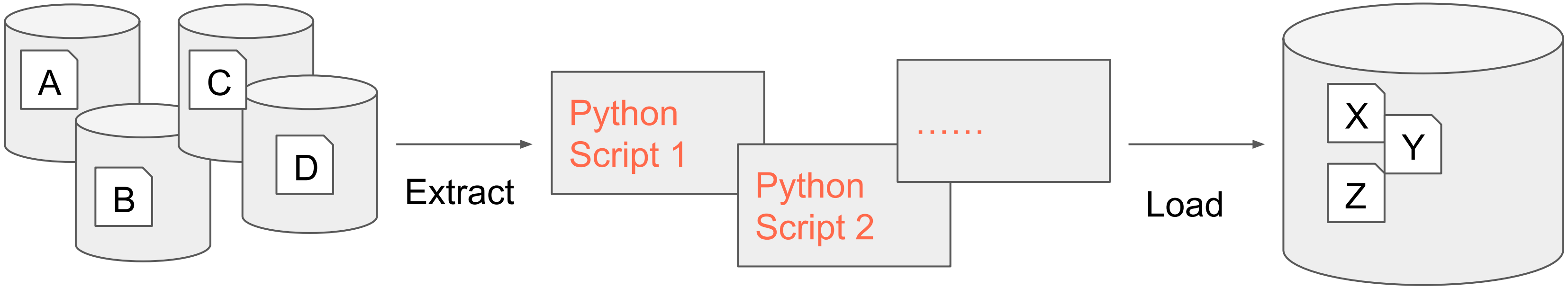
What is transformation

- Transform the shape of the data in operational databases to fit the shape of the data in data warehouse
  - Operational databases only record latest information
  - Data warehouse stores historical information for the purposes of analysis

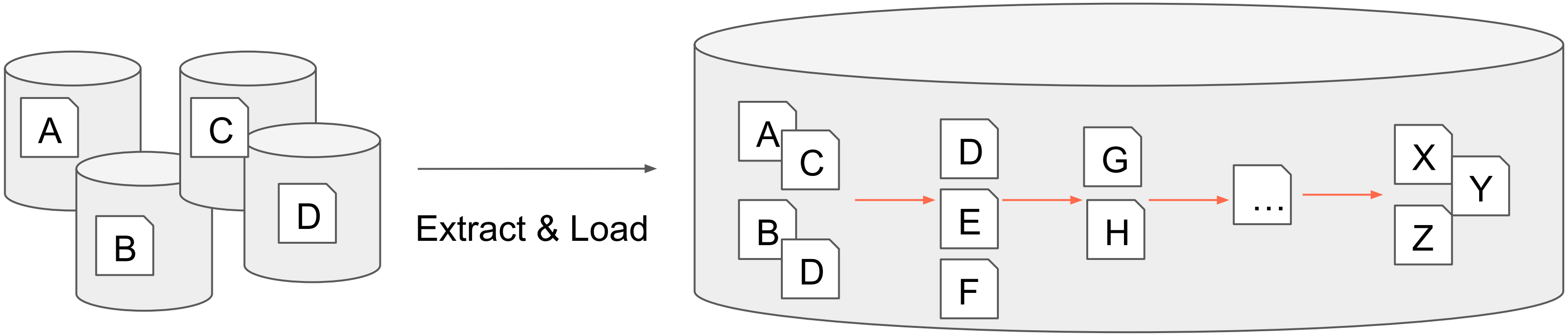
- Why transformation is complicated
  - Database and datawarehouse serves different purposes → different schema designs → different shape of the data
  - Multiple sources, but only one destination

What is transformation

ETL



ELT



What is dbt

- A tool extremely good at transforming data that’s already loaded into your warehouse (ELT)
    - How dbt make it ?
        - SQL + Jinja + Yaml

- dbt is designed for ease of use in **data engineering**: for when you need to develop a data pipeline.

  — dbtvault

- Tools that empower **analysts** to own the entire analytics engineering workflow

  — dbt’s mission statement

What is dbt

SQL script

```
daily_revenue_by_product.sql
SELECT
product_id, order_date, SUM(amount) as
ttl_amount
FROM orders
GROUP BY produtct_id, date
```

- Copy results to excel
- Captured by visualization tools

What is dbt

SQL script

```
daily_revenue_by_product.sql
SELECT
product_id, order_date, SUM(amount) as
ttl_amount
FROM orders
GROUP BY product_id, date
```

- Copy results to excel
- Captured by visualization tools

dbt SQL script

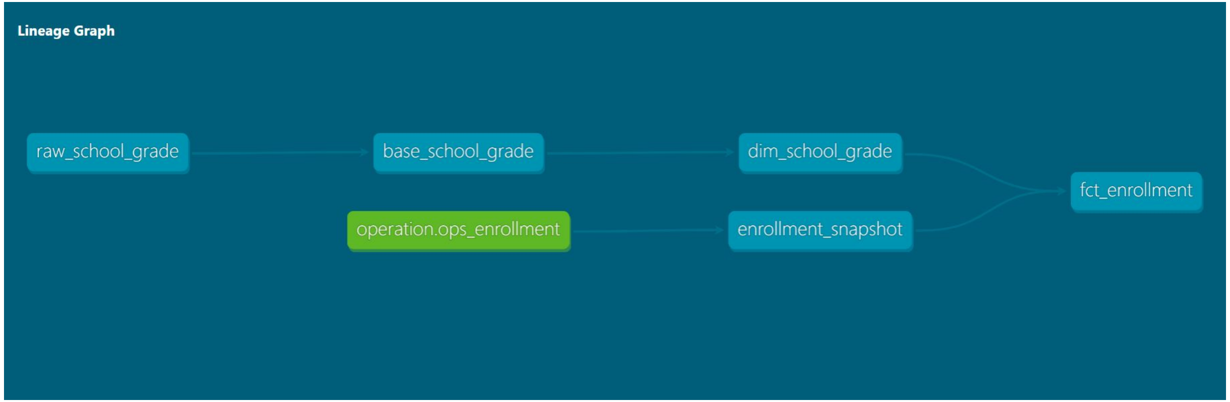
```
daily_revenue_by_product.sql
SELECT
product_id, order_date, SUM(amount) as
ttl_amount
FROM {{ ref('orders') }}
GROUP BY product_id, date
```

- Write a table/model into data warehouse
- No DDL / DML
    - DDL: create, alter, drop, truncate
    - DML: update, insert

What does dbt provide ?

- Transformation
- Data quality (unit test)
- Data lineage

```
models:
  - name: table1
    columns:
      - name: enrollment_id
        description: "Primary Key"
        tests:
          - unique
          - not_null
```



What does dbt provide ?

- Version control and CI/CD
- Macros
- Exposure
- Metrics
- Freshness
- Latest: python-based transformation

dbt does more than transformation

Agenda

What is dbt ?

- What is transformation
- What is dbt
- What does dbt provide

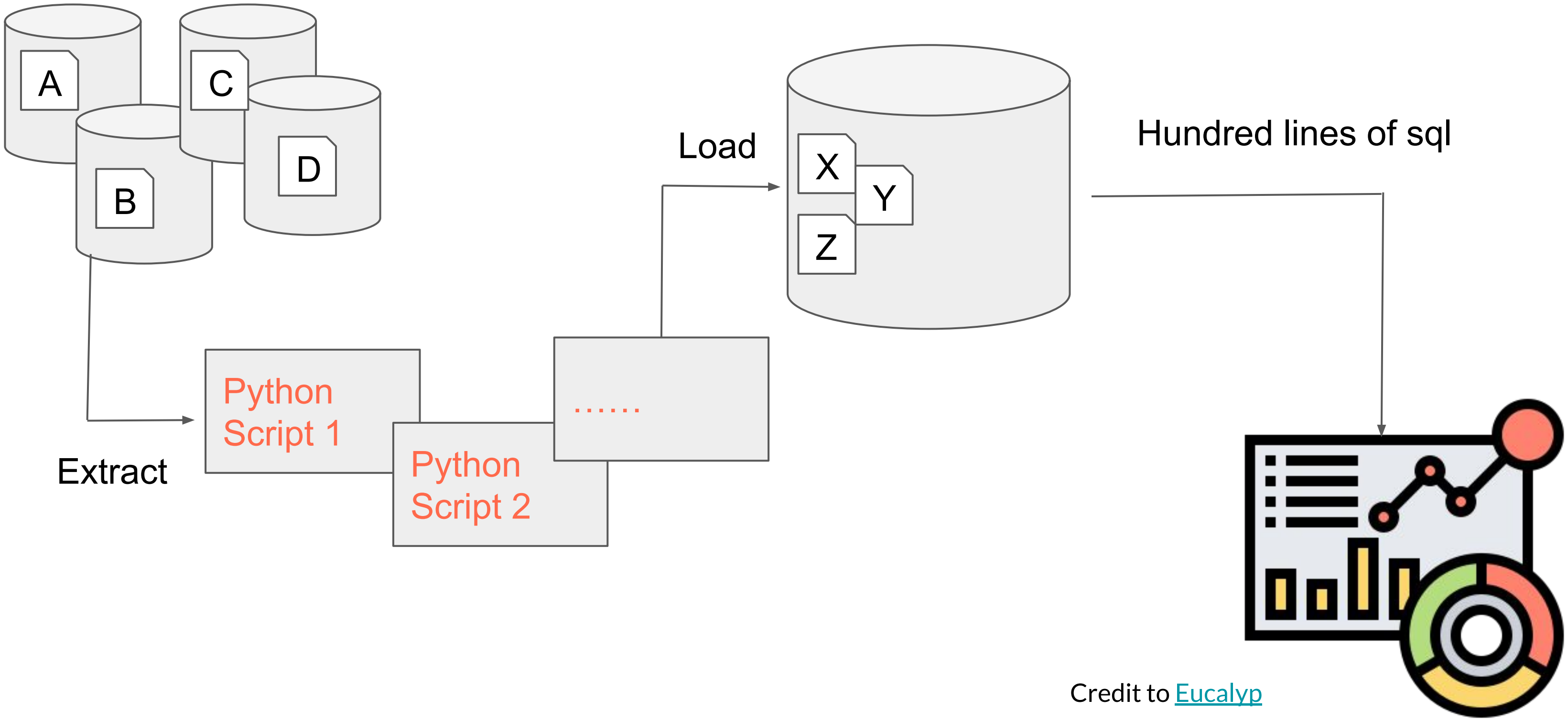
Why people love dbt

- Why organizations embrace dbt
- Why Data engineer love dbt
- Why I love dbt

How to get started ?

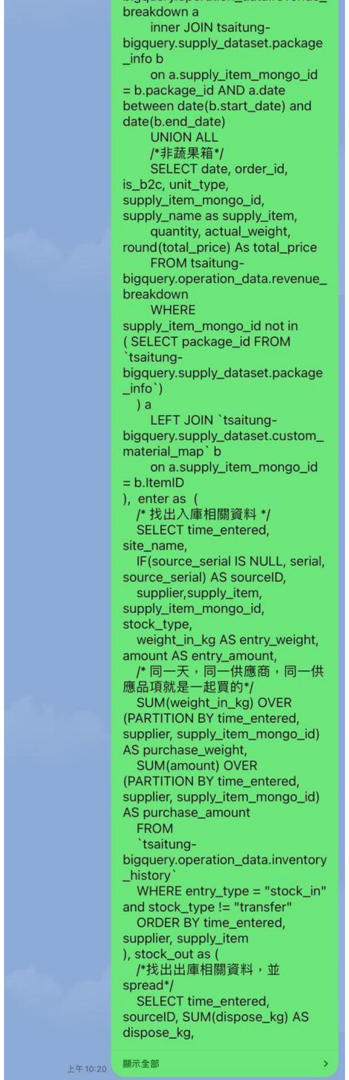
- Prerequisites
- Installation and guide
- Become a pro

Before using dbt



Credit to Eucalyp

Before using dbt

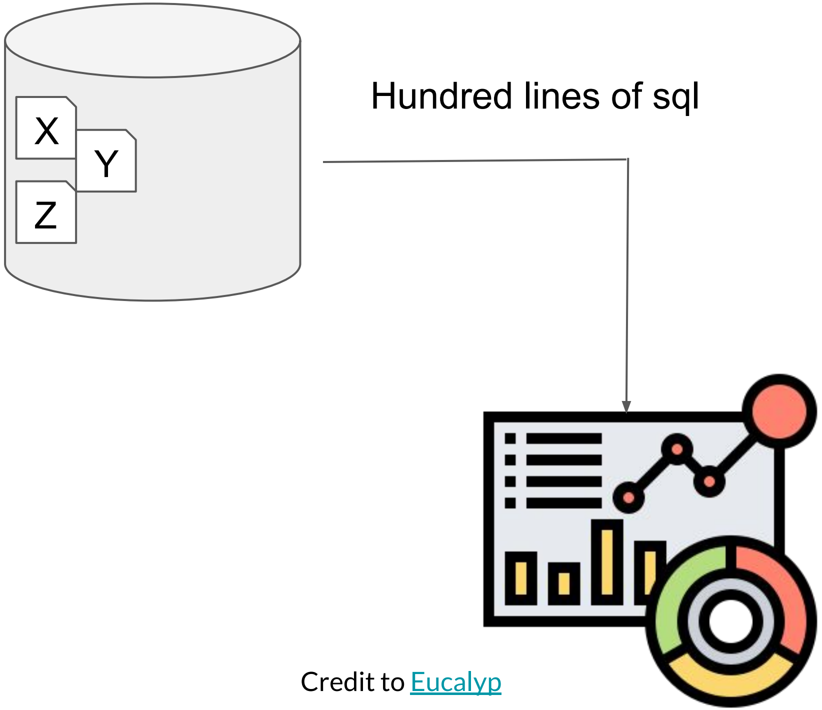


```
breakdown a
      inner JOIN tsaitung-
bigquery.supply_dataset.package
_info b
      on a.supply_item_mongo_id
= b.package_id AND a.date
between date(b.start_date) and
date(b.end_date)
      UNION ALL
      /*非蔬果箱*/
      SELECT date, order_id,
is_b2c, unit_type,
supply_item_mongo_id,
supply_name as supply_item,
      quantity, actual_weight,
round(total_price) As total_price
      FROM tsaitung-
bigquery.operation_data.revenue_
breakdown
      WHERE
supply_item_mongo_id not in
( SELECT package_id FROM
`tsaitung-
bigquery.supply_dataset.package
_info`)
   ) a
      LEFT JOIN `tsaitung-
bigquery.supply_dataset.custom_
material_map` b
      on a.supply_item_mongo_id
= b.ItemID
),  enter as  (
   /* 找出入庫相關資料 */
   SELECT time_entered,
site_name,
   IF(source_serial IS NULL, serial,
source_serial) AS sourceID,
   supplier,supply_item,
supply_item_mongo_id,
stock_type,
   weight_in_kg AS entry_weight,
amount AS entry_amount,
   /* 同一天，同一供應商，同一供
應品項就是一起買的*/
   SUM(weight_in_kg) OVER
(PARTITION BY time_entered,
supplier, supply_item_mongo_id)
AS purchase_weight,
   SUM(amount) OVER
(PARTITION BY time_entered,
supplier, supply_item_mongo_id)
AS purchase_amount
   FROM
   `tsaitung-
bigquery.operation_data.inventory
_history`
   WHERE entry_type = "stock_in"
and stock_type != "transfer"
   ORDER BY time_entered,
supplier, supply_item
), stock_out as (
   /*找出出庫相關資料，並
spread*/
   SELECT time_entered,
sourceID, SUM(dispose_kg) AS
dispose_kg,
```

Before using dbt

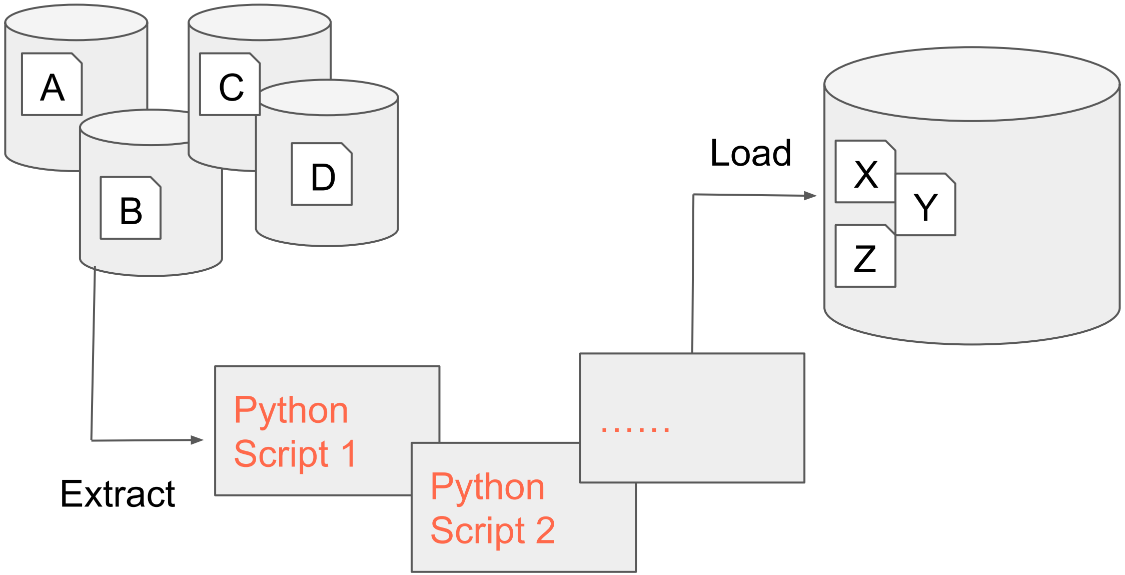
Credit to Eucalyp

Before using dbt

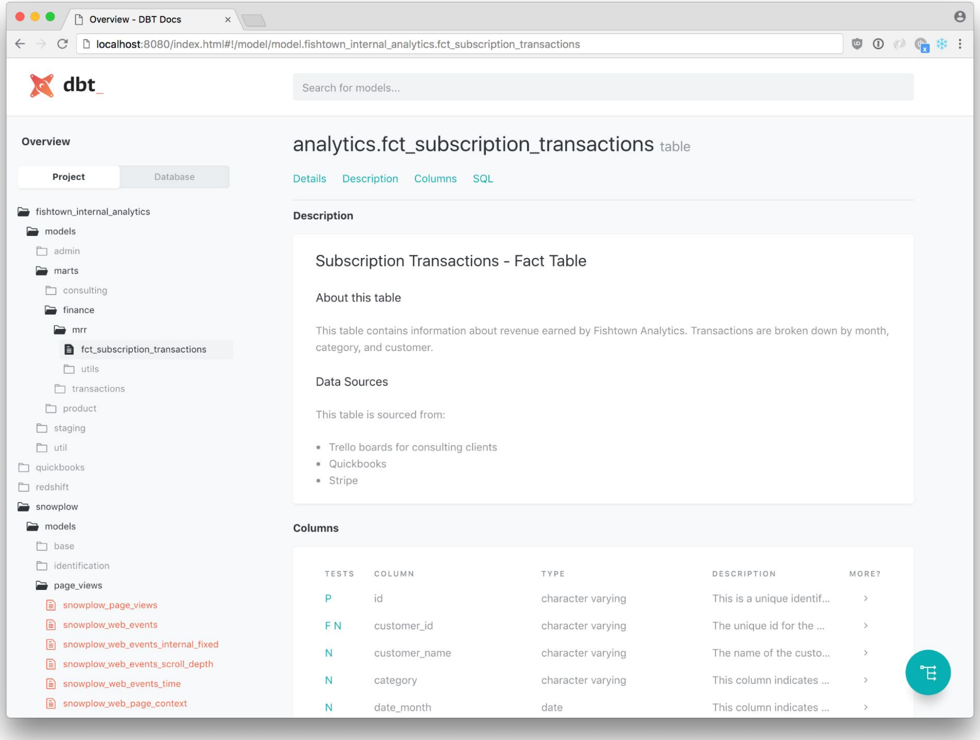


Credit to Eucalyp

Before using dbt



After using dbt



Overview - DBT Docs
localhost:8080/index.html#!/model/model.fishtown_internal_analytics.fct_subscription_transactions
dbt_
Search for models...
Overview
Project
Database

- fishtown_internal_analytics
  - models
    - admin
    - marts
      - consulting
      - finance
        - mrr
          - fct_subscription_transactions
          - utils
        - transactions
      - product
    - staging
    - util
- quickbooks
- redshift
- snowplow
  - models
    - base
    - identification
    - page_views
      - snowplow_page_views
      - snowplow_web_events
      - snowplow_web_events_internal_fixed
      - snowplow_web_events_scroll_depth
      - snowplow_web_events_time
      - snowplow_web_page_context

analytics.fct_subscription_transactions table

Details Description Columns SQL

Description

Subscription Transactions - Fact Table

About this table

This table contains information about revenue earned by Fishtown Analytics. Transactions are broken down by month, category, and customer.

Data Sources

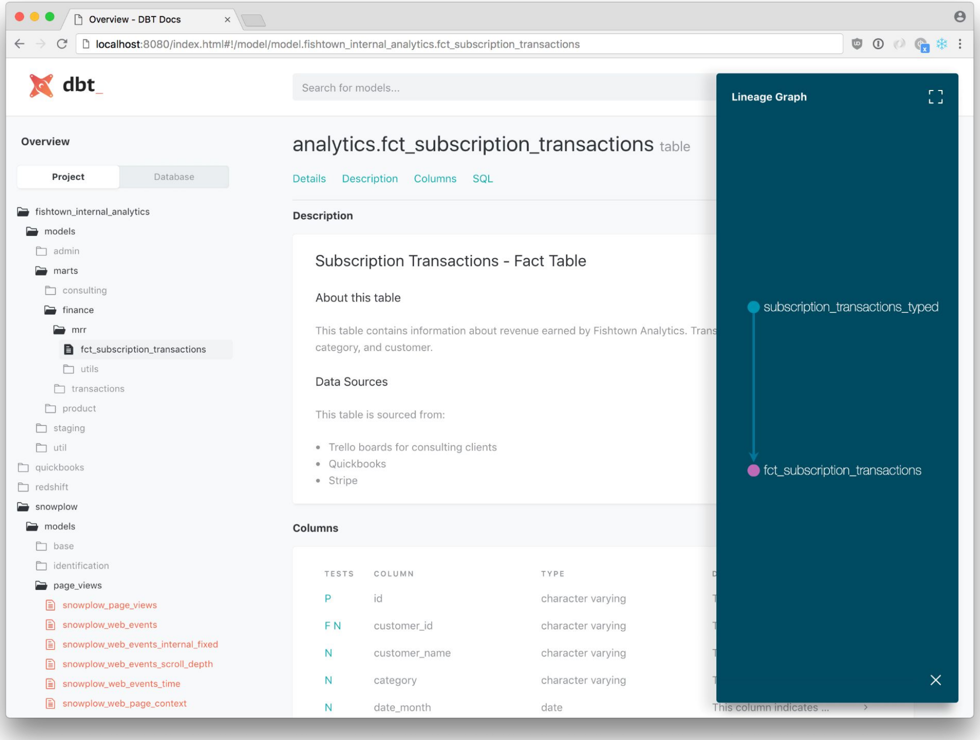
This table is sourced from:

- Trello boards for consulting clients
- Quickbooks
- Stripe

Columns

TESTS	COLUMN	TYPE	DESCRIPTION	MORE?
P	id	character varying	This is a unique identif...	›
F N	customer_id	character varying	The unique id for the ...	›
N	customer_name	character varying	The name of the custo...	›
N	category	character varying	This column indicates ...	›
N	date_month	date	This column indicates ...	›

After using dbt


Overview - DBT Docs
localhost:8080/index.html#!/model/model.fishtown_internal_analytics.fct_subscription_transactions
dbt_
Search for models...
Overview
Project
Database
fishtown_internal_analytics
models
admin
marts
consulting
finance
mrr
fct_subscription_transactions
utils
transactions
product
staging
util
quickbooks
redshift
snowplow
models
base
identification
page_views
snowplow_page_views
snowplow_web_events
snowplow_web_events_internal_fixed
snowplow_web_events_scroll_depth
snowplow_web_events_time
snowplow_web_page_context
Lineage Graph
subscription_transactions_typed
fct_subscription_transactions

analytics.fct_subscription_transactions table

Details Description Columns SQL

Description

Subscription Transactions - Fact Table

About this table

This table contains information about revenue earned by Fishtown Analytics. Trans

category, and customer.

Data Sources

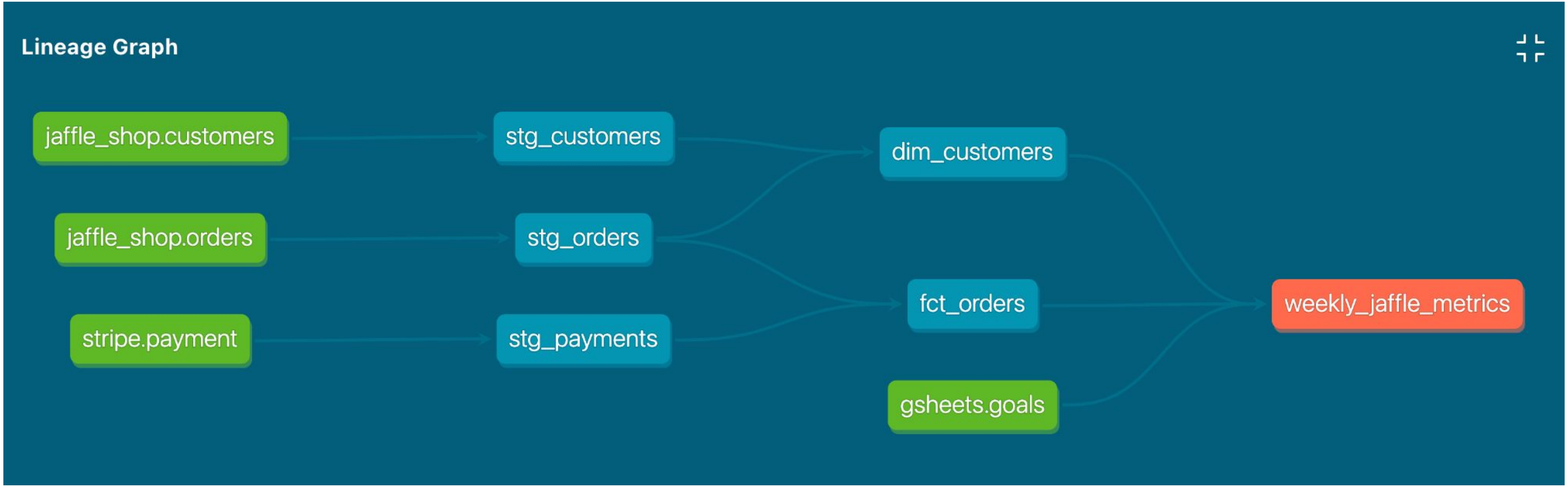
This table is sourced from:

- Trello boards for consulting clients
- Quickbooks
- Stripe

Columns

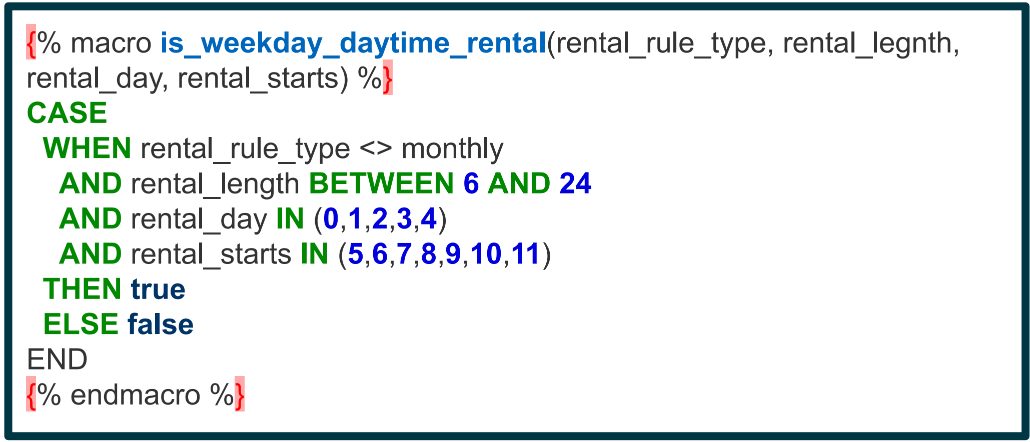
TESTS	COLUMN	TYPE	D
P	id	character varying	T
F N	customer_id	character varying	T
N	customer_name	character varying	T
N	category	character varying	T
N	date_month	date	This column indicates ... ›

After using dbt



After using dbt

- Transparency
- Consistency
  - Reference from same upstream model
  - Macros
- Ease of use
  - SQL is intuitive



```
{% macro is_weekday_daytime_rental(rental_rule_type, rental_legngth,
rental_day, rental_starts) %}
CASE
  WHEN rental_rule_type <> monthly
    AND rental_length BETWEEN 6 AND 24
    AND rental_day IN (0,1,2,3,4)
    AND rental_starts IN (5,6,7,8,9,10,11)
  THEN true
  ELSE false
END
{% endmacro %}
```

Agenda

What is dbt ?

- What is transformation
- What is dbt
- What does dbt provide

Why people love dbt

- Why organizations embrace dbt
- Why Data engineer love dbt
- Why I love dbt

How to get started ?

- Prerequisites
- Installation and guide
- Become a pro

Why data engineers love dbt

- Data engineers can do all the functions that dbt have.
- However, it’s quite bothering, repetitive, and time-consuming it they need to do most the transformation and maintenance themselves
- It can free data engineers’ time to do more meaningful things (i.e. develop new tools)

Transformation – when a new table created

Without dbt

```
create target_table (
  id varchar(10),
  col2 int,
  date timestamp
)

insert into target_table
select  id, col2, date from upstream_table
```

dbt script

```
{{ config( model = 'incremental',
                 unique_key = 'id',
                 on_schema_change =
'sync_all_columns')
}}

select  id, col2, date
from {{ ref('upstream_table') }}
{{% if is_incremental() %}}
WHERE date > (select max(date) from {{ this }})
{{% endif %}}
```

Transformation– insert new rows into an existing table

Without dbt

```
insert into target_table
select  id, col2, date from upstream_table
WHERE date >
        (select max(date) from target_table)
```

dbt script

```
{{ config( model = 'incremental',
           unique_key = 'id',
           on_schema_change =
'sync_all_columns')
}}

select  id, col2, date
from {{ ref('upstream_table') }}
{{% if is_incremental() %}}
WHERE date > (select max(date) from {{ this }})
{{% endif %}}
```

Transformation– add a column to an existing table

Without dbt

```
alter target_table
add col4 int ;

select  id, col2, date, col4
from upstream_table
WHERE date > (select max(date) from target_table)
```

dbt script

```
{{ config( model = 'incremental',
           unique_key = 'id',
           on_schema_change =
'sync_all_columns')
}}

select  id, col2, date,col4
from {{ ref('upstream_table') }}
{{% if is_incremental() %}}
WHERE date > (select max(date) from {{ this }})
{{% endif %}}
```

Transformation– other cases

Without dbt

- Update an existing row
  → Another Script / DDL /DML
- Change a table to a view
- Remove a column
- More other cases

dbt script

```
{{ config( model = 'incremental',
                unique_key = 'id',
                on_schema_change =
'sync_all_columns')
}}

select  id, col2, col3
from {{ ref('upstream_table') }}
{{% if is_incremental() %}}
WHERE col3 > (select max(col3) from {{ this }})
{{% endif %}}
```

Data lineage graph

Without dbt

Sorry, I don’t how to create one on

(but probabl

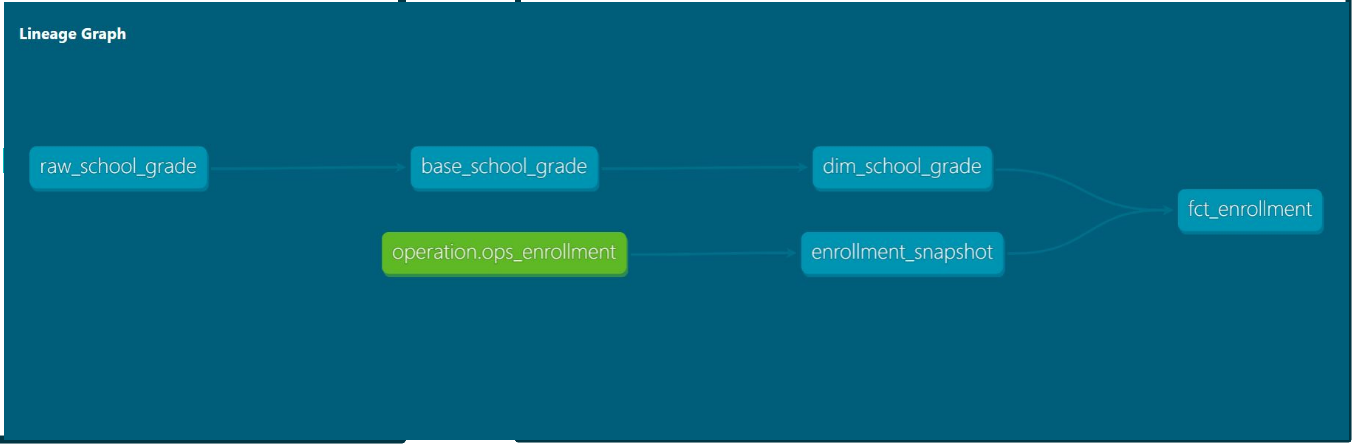
dbt

base_school_grade.sql

```
select * from {{ ref('raw_school_grade') }}
```

dim_school_grade.sql

```
select * from {{ ref('base_school_grade') }}
```



Data quality

Without dbt

```sql
select enrollment_id, count(*) as int
from table1 group by 1 having cnt > 1 ;

select enrollment_id from table1
WHERE enrollment_id is null ;
```

dbt

```yaml
models:
  - name: table1
    columns:
      - name: enrollment_id
        description: "Primary Key"
        tests:
          - unique
          - not_null
```

Agenda

What is dbt ?

- What is transformation
- What is dbt
- What does dbt provide

Why people love dbt

- Why organizations embrace dbt
- Why Data engineer love dbt
- Why I love dbt

How to get started ?

- Prerequisites
- Installation and guide
- Become a pro

Why I love dbt

As a data analyst,

- Transparency
- Consistency
- Ease of use
  - Leverage others’ SQL code and model
- Jinja

Jinja – deal with repetitive SQL

For example: pivot table

```
vars:
  school_grades: ['學前一', '學前二', '學前三','幼小', '幼中', '幼大', '1年級', '2年級', '3年級', '4年級', '5年級', '6年級' ]

{% set gs = var('school_grades') %}
With pivot_result as (
        SELECT
        material_level, material_sequence, subject, first_enrollment_date, first_enrolled_center,
        concat(material_level, material_sequence, subject, first_enrollment_date, first_enrolled_center) AS uni_key,
        {% for g in gs %}
        SUM(case when school_grade = '{{ g }}' then 1 else 0 end) as "{{g}}"
        {% if not loop.last %}, {% endif %}
        {% endfor %}
        FROM result
        group by material_level, material_sequence, subject, first_enrollment_date, first_enrolled_center
)


SELECT * FROM pivot_result
```

New skill, new knowledge, and new world

Data analyst usually equipped with

- (Solid) SQL Skill
- (Advanced) statistics knowledge
- (Strong) Empathy and Curiosity

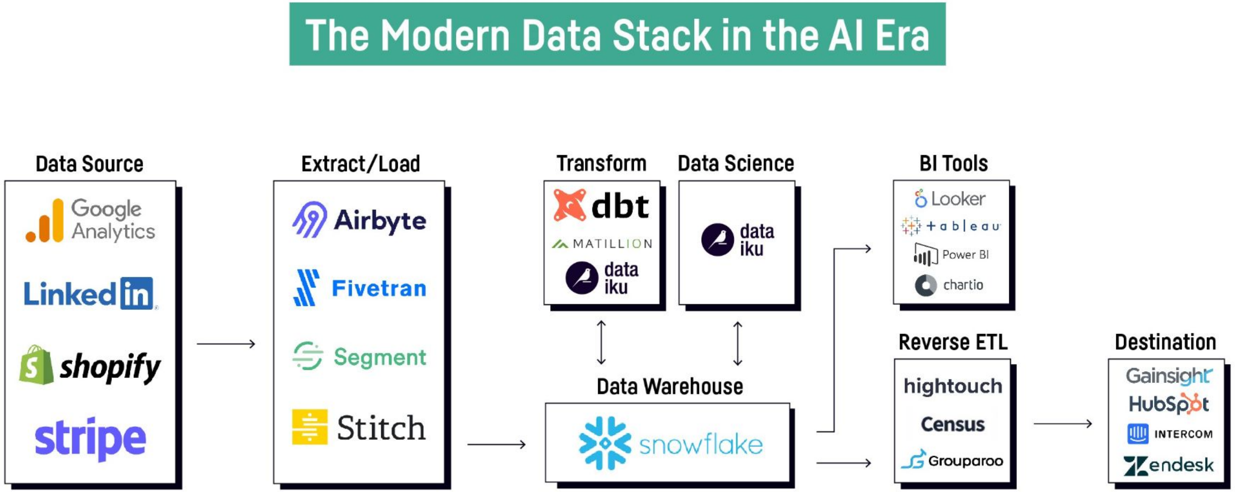
However, the career seems limited.

After knowing dbt, I learnt about

- How to do transformation
- Batch processing and scheduling
- SWE best practice

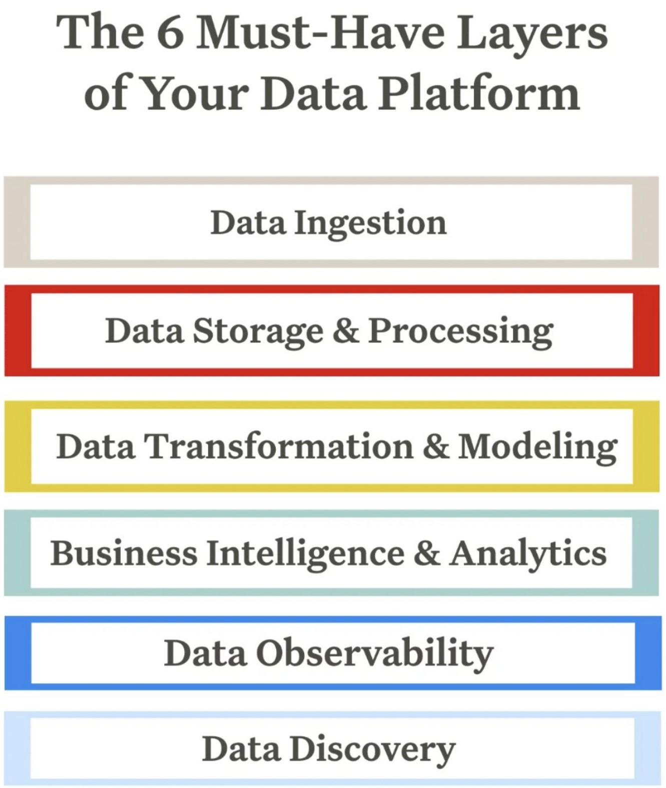
And more importantly, I start to have a big picture of a data team / data stack

New skill, new knowledge, and new world



Source: Dataiku

New skill, new knowledge, and new world



The 6 Must-Have Layers of Your Data Platform

Data Ingestion

Data Storage & Processing

Data Transformation & Modeling

Business Intelligence & Analytics

Data Observability

Data Discovery

Source: MonteCarlo

Agenda

What is dbt ?

- What is transformation
- What is dbt
- What does dbt provide

Why people love dbt

- Why organizations embrace dbt
- Why Data engineer love dbt
- Why I love dbt

How to get started ?

- Prerequisites
- Installation and guide
- Become a pro

Prerequisites

- Python
    - Because dbt is python-based
- A data warehouse
    - Of course, BigQuery here
- Know SQL
    - All you need to know is SELECT
- Ubuntu Environment
    - Optional, but nice to have
    - Windows works too (but don't ask me how to install database on windows)

How to get started

- Follow dbt’s installation guide

- Watch dbt’s tutorial video
  - After watching, just start to write your first model (.sql script)

- Navigate through dbt’s website
  - Best practice
  - Discourse
  - Join dbt slack community and local-taipei channel

Become a pro

- Think carefully, instead of just jumping in and writing a model.
  - When pipelines become complicated, maintainability and scalability is important.
  - What should the target schema/table should look like ?
    - One big table ?
    - Star schema ?
- Read books about data modeling
  - The data warehouse toolkit
- Follow data experts to get the latest news of modern data stack
  - Chad Anderson
  - Christian Kaul

Thank you!

11/19 Sat,

Reverse ETL and Morden Data Stack

Taipei dbt Meetup Group
#local-taipei Slack Channel

RSVP

dbt