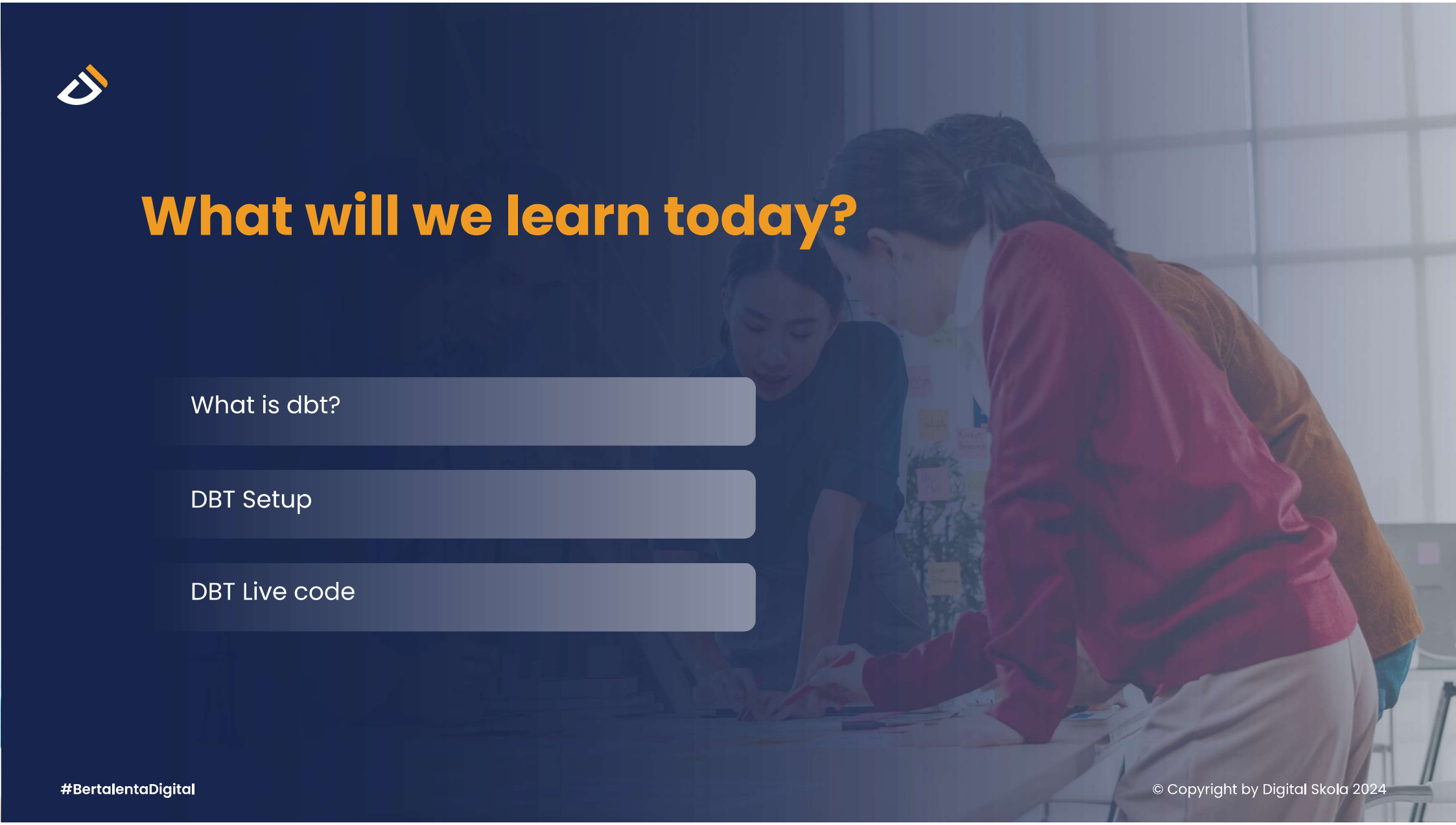
Data Build Tools (DBT)

What will we learn today?

- What is dbt?
- DBT Setup
- DBT Live code



#BertalentaDigital

What is DBT?

We start with Excel files

	A	B	C
1	Item	Qty.	Delivery
2	Sweets	20	Delivered
3	Biscuits	100	In transit
4	Ice-cream	50	Delivered
5	Juice	95	Delivered

Products1 | Produ

	A	B	C
1	Item	Qty.	Delivery
2	Cakes	120	Delivered
3	Croissants	70	In transit
4	Appe pies	75	Past due
5	Doughnuts	200	Delivered

... Products2 | Produ

	A	B	C
1	Item	Qty.	Delivery
2	Strawberry	120	Delivered
3	Bilberry	70	In transit
4	Raspberry	85	Past due
5	Blackberry	110	In transit

... Products3

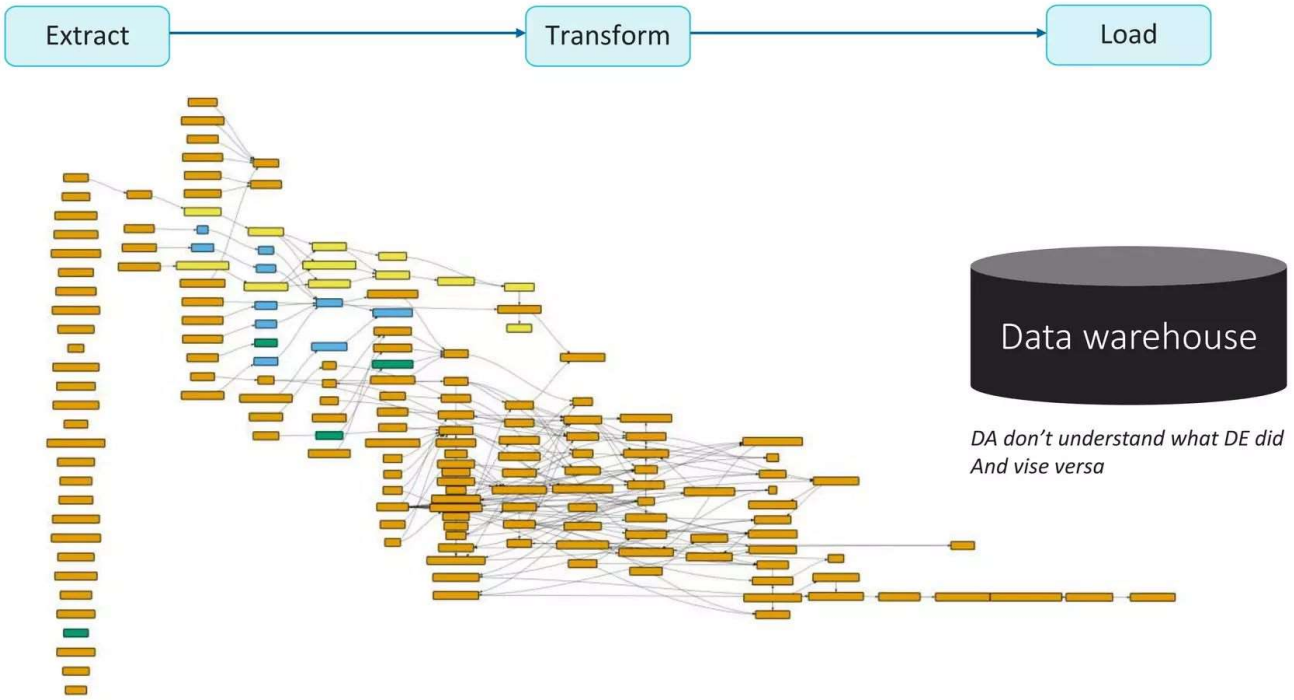
	A	B	C
1	Item	Qty.	Delivery
2	Sweets	20	Delivered
3	Biscuits	100	In transit
4	Ice-cream	50	Delivered
5	Juice	95	Delivered
6	Lollipops	90	Past due
7	Cakes	120	Delivered
8	Croissants	70	In transit
9	Appe pies	75	Past due
10	Doughnuts	200	Delivered
11	Pastry	150	In transit
12	Strawberry	120	Delivered
13	Bilberry	70	In transit
14	Raspberry	85	Past due
15	Blackberry	110	In transit
16	Honeyberry	90	Delivered

... Summary_sheet

<epam>

Background

We moved to Datawarehouse



<epam>

Lead time at least 2 weeks

Background

DE challenges

```sql
WITH order_summary AS (
  SELECT
    DATE_TRUNC('month', o.order_date) AS month,
    o.order_id,
    SUM(oi.price * oi.quantity) AS revenue
  FROM
    orders o
  JOIN
    order_items oi ON o.order_id = oi.order_id
  GROUP BY
    DATE_TRUNC('month', o.order_date),
    o.order_id
), bill_summary AS (
  SELECT
    month,
    COUNT(DISTINCT order_id) AS total_bills,
    SUM(revenue) AS total_revenue,
    AVG(revenue) AS avg_bill_value
  FROM
    order_summary
  GROUP BY
    month
)
SELECT
  month,
  total_bills,
  total_revenue,
  avg_bill_value
FROM
  bill_summary
ORDER BY
  month ASC;
```

Readability
- How to read and understand this query?
- Where to start?

Accessibility
- How to verify the output?
- Can we break the script into smaller pieces for testing?

Collaboration
- How to reuse this query for other analysis?
- How to onboard new members?
- How to explain if there're 100 tables?

Scripting
- How to reuse this query for other analysis?
- How to manage model versions?

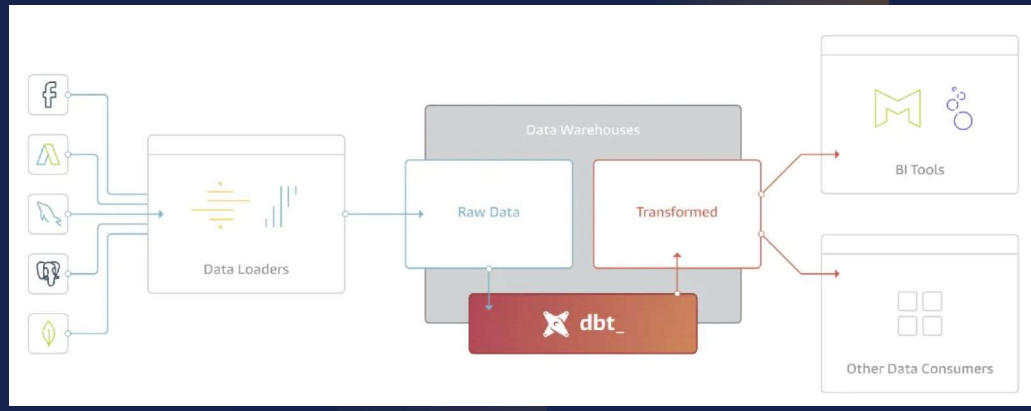
Background

Data Build Tools (DBT)

dbt is a command line tool that enables data analysts and engineers to transform data in their warehouses more effectively. Today, dbt has ~850 companies using it in production, including companies like Casper, Seatgeek, and Wistia.

Data Build Tools (DBT)

dbt fits nicely into the modern BI stack, coupling with products like Stitch, Fivetran, Redshift, Snowflake, BigQuery, Looker, and Mode. Here's how the pieces fit together:



Data Build Tools (DBT)

dbt is the T in ELT. It doesn't extract or load data, but it's extremely good at transforming data that's already loaded into your warehouse. This "transform after load" architecture is becoming known as ELT (extract, load, transform).

ELT has become commonplace because of the power of modern analytic databases. Data warehouses like Redshift, Snowflake, and BigQuery are extremely performant *and* very scalable such that at this point most data transformation use cases can be much more effectively handled in-database rather than in some external processing layer. Add to this the separation of compute and storage and there are decreasingly few reasons to want to execute your data transformation jobs elsewhere.

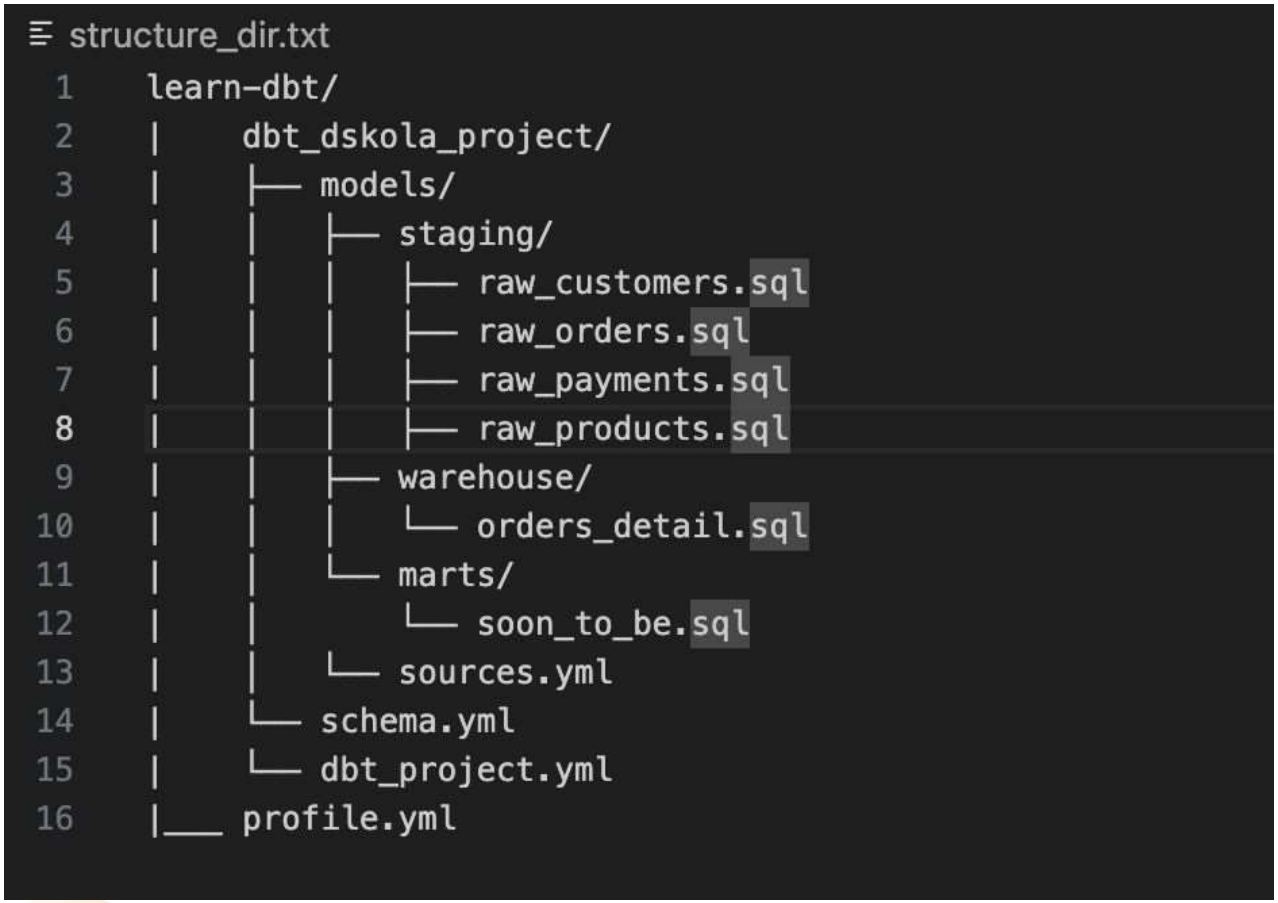
dbt is a tool to help you write and execute the data transformation jobs that run inside your warehouse. **dbt's only function is to take code, compile it to SQL, and then run against your database.**

DBT Setup

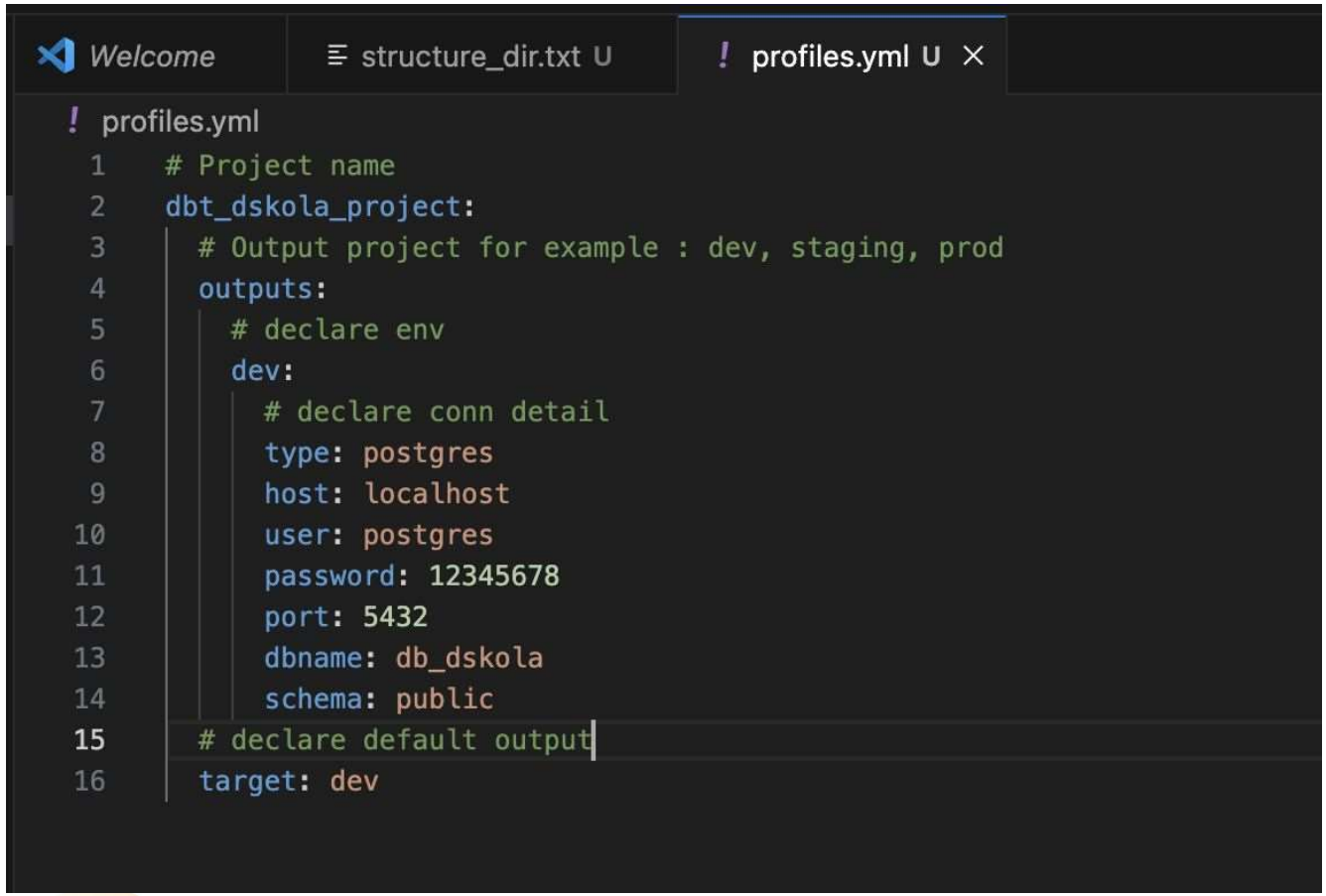
DBT Setup

https://docs.getdbt.com/docs/core/pip-install

DBT Live Code



DBT Structure Dir



DBT Profiles explanation

DBT Live Code

Important command :

dbt debug --profiles-dir ./ --project-dir dbt_dskola_project

dbt run --profiles-dir ./ --project-dir dbt_dskola_project

dbt init <project_name>

dbt docs generate --profiles-dir ./ --project-dir dbt_dskola_project

dbt docs serve --profiles-dir ./ --project-dir dbt_dskola_project

DBT Live Code

https://gitlab.com/farhansmg/learn_dbt

Project 2

Source data :
https://github.com/graphql-compose/graphql-compose-examples/tree/master/examples/northwind/data/csv

Dari source data diatas, buatlah :
1. Raw data nya masukkan ke postgre/snowflake (menggunakan python)
2. Buatkan datawarehouse nya

Lakukan semuanya didalam dbt

Project 2

Setelah dimasukkan buatkan data mart berikut :
Sebagai marketing leader, saya ingin dashboard report berikut:
(gross revenue = (harga - (harga * diskon))* jumlah barang)

1. Table supplier (company_name) gross revenue tiap bulan
   (datamart_monthly_supplier_gross_revenue)
2. tabel kategori produk paling banyak terjual tiap bulan
   (datamart_monthly_category_sold)
3. tabel best employee (employee_name) berdasarkan total gross revenue yang dihasilkan dalam satu bulan (datamart_monthly_best_employee)

Project 2

Bonus :

1. Running all of them (python, postgre, dbt) menggunakan docker (pisahkan image nya)

Well done :

1. Bisa menggunakan snowflake / bigquery as datawarehouse

Project 2

Kumpulkan :

1. Dalam 1 folder (contohnya seperti di git https://gitlab.com/farhansmg/learn_dbt)
2. Select * from data mart hasil jawaban kalian, limit 10 → screenshoot → masukkan ss tersebut ke folder yang sama seperti no.1

DigitalSkola
Uncover The World of Digital Skills with Us

Thank You

#BertalentaDigital | digitalskola.com