

Департамент образования и науки города Москвы

**Государственное автономное образовательное
учреждение высшего образования города Москвы
«Московский городской педагогический университет»**

Институт цифрового образования

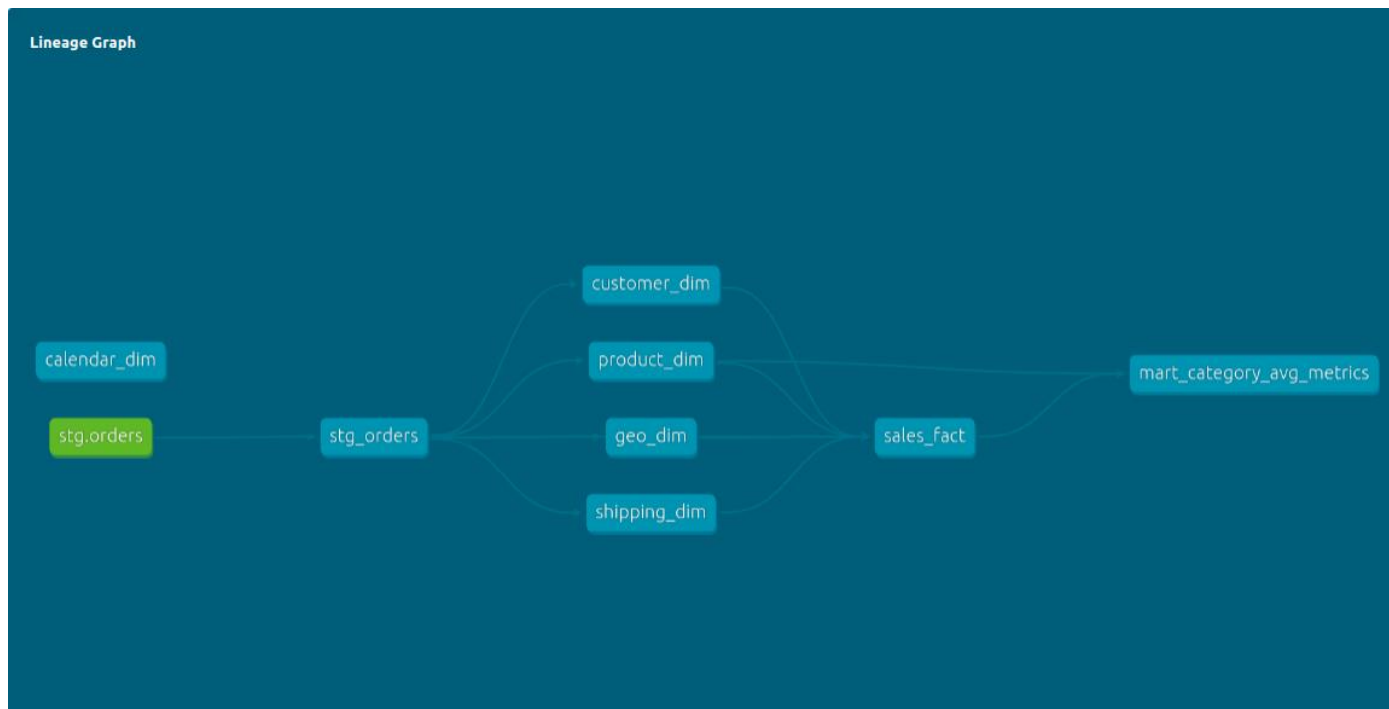
**Практическая работа 2.1 по дисциплине
«Платформы Data Engineering»**

Выполнил: студент БД-251м

Быков Владимир

Москва 2025 г.

1. Архитектура DWH



2. Ключевые фрагменты кода

- Код модели `stg_orders.sql`

SELECT

-- Приводим все к нижнему регистру для консистентности в dbt

"order_id",

("order_date")::date as order_date,

("ship_date")::date as ship_date,

"ship_mode",

"customer_id",

"customer_name",

"segment",

"country",

"city",

"state",

-- Исправляем проблему с Burlington прямо здесь, один раз и навсегда

CASE

WHEN "city" = 'Burlington' AND "postal_code" IS NULL THEN '05401'

ELSE "postal_code"

END as postal_code,

```

"region",
"product_id",
"category",
"subcategory" as sub_category, -- переименовываем для соответствия
"product_name",
"sales",
"quantity",
"discount",
"profit"
FROM {{ source('stg', 'orders') }}

```

- **Код модели sales_fact.sql**

-- Создает таблицу фактов, объединяя все измерения

```
SELECT
```

```
-- Суррогатные ключи из измерений
```

```
cd.cust_id,
```

```
pd.prod_id,
```

```
sd.ship_id,
```

```
gd.geo_id,
```

```
-- Ключи для календаря
```

```
to_char(o.order_date, 'yyyymmdd')::int AS order_date_id,
```

```
to_char(o.ship_date, 'yyyymmdd')::int AS ship_date_id,
```

```
-- Бизнес-ключ и метрики
```

```
o.order_id,
```

```
o.sales,
```

```
o.profit,
```

```
o.quantity,
```

```
o.discount
```

```
FROM {{ ref('stg_orders') }} AS o
```

```
LEFT JOIN {{ ref('customer_dim') }} AS cd ON o.customer_id = cd.customer_id
```

```
LEFT JOIN {{ ref('product_dim') }} AS pd ON o.product_id = pd.product_id
```

```
LEFT JOIN {{ ref('shipping_dim') }} AS sd ON o.ship_mode = sd.ship_mode
```

```
LEFT JOIN {{ ref('geo_dim') }} AS gd ON o.postal_code = gd.postal_code AND o.city =
gd.city AND o.state = gd.state
```

- **Код индивидуальной mart-модели mart_category_avg_metrics.sql**

--Эффективность товарных категорий. Рассчитать средний чек и среднюю

--прибыль на один заказ для каждой категории товаров.

```
SELECT
```

```
    p.category,
```

```
    COUNT(DISTINCT f.order_id) as order_count,
```

```
    ROUND(AVG(f.sales), 2) as avg_sales_per_order,
```

```
    ROUND(AVG(f.profit), 2) as avg_profit_per_order
```

```
FROM {{ ref('sales_fact') }} AS f
```

```
LEFT JOIN {{ ref('product_dim') }} AS p ON f.prod_id = p.prod_id
```

```
GROUP BY p.category
```

```
ORDER BY avg_sales_per_order DESC
```

- **Файл schema.yml с тестами для всех моделей**

```
version: 2
```

```
models:
```

```
- name: shipping_dim
```

```
  columns:
```

```
    - name: ship_id
```

```
    tests:
```

```
      - unique
```

```
      - not_null
```

```
- name: customer_dim
```

```
  columns:
```

```
    - name: cust_id
```

```
    tests:
```

```
      - unique
```

```
      - not_null
```

```
- name: geo_dim
```

columns:

- name: geo_id

tests:

- unique

- not_null

- name: mart_category_avg_metrics

columns:

- name: category

tests:

- not_null

- name: avg_sales_per_order

tests:

- not_null

- name: product_dim

columns:

- name: prod_id

tests:

- unique

- not_null

- name: sales_fact

columns:

- name: cust_id

tests:

- relationships:

arguments:

to: ref('customer_dim')

field: cust_id

3. Результаты

- Скриншот успешного выполнения dbt run и dbt test для проекта student_dwh

```

• (dbt-env) dev@dev-vm:~/Downloads/pde_magistr/superstore_dwh$ dbt run --select mart_category_avg_metrics
08:41:32 Running with dbt=1.10.11
08:41:32 Registered adapter: postgres=1.9.1
08:41:33 Found 8 models, 11 data tests, 1 source, 435 macros
08:41:33
08:41:33 Concurrency: 4 threads (target='dev')
08:41:33
08:41:33 1 of 1 START sql table model dw_test.mart_category_avg_metrics ..... [RUN]
08:41:34 1 of 1 OK created sql table model dw_test.mart_category_avg_metrics ..... [SELECT 3 in 0.74
s]
08:41:34
08:41:34 Finished running 1 table model in 0 hours 0 minutes and 0.89 seconds (0.89s).
08:41:34
08:41:34 Completed successfully
08:41:34
08:41:34 Done. PASS=1 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=1

• (dbt-env) dev@dev-vm:~/Downloads/pde_magistr/superstore_dwh$ dbt test --select mart_category_avg_metrics
08:43:29 Running with dbt=1.10.11
08:43:30 Registered adapter: postgres=1.9.1
08:43:30 Found 8 models, 11 data tests, 1 source, 435 macros
08:43:30
08:43:30 Concurrency: 4 threads (target='dev')
08:43:30
08:43:30 1 of 2 START test not_null_mart_category_avg_metrics_avg_sales_per_order ..... [RUN]
08:43:30 2 of 2 START test not_null_mart_category_avg_metrics_category ..... [RUN]
08:43:30 1 of 2 PASS not_null_mart_category_avg_metrics_avg_sales_per_order ..... [PASS in 0.10s]
08:43:30 2 of 2 PASS not_null_mart_category_avg_metrics_category ..... [PASS in 0.10s]
08:43:30
08:43:30 Finished running 2 data tests in 0 hours 0 minutes and 0.26 seconds (0.26s).
08:43:30
08:43:30 Completed successfully
08:43:30
08:43:30 Done. PASS=2 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=2
○ (dbt-env) dev@dev-vm:~/Downloads/pde_magistr/superstore_dwh$

```

• Скриншот с данными из индивидуальной mart-модели

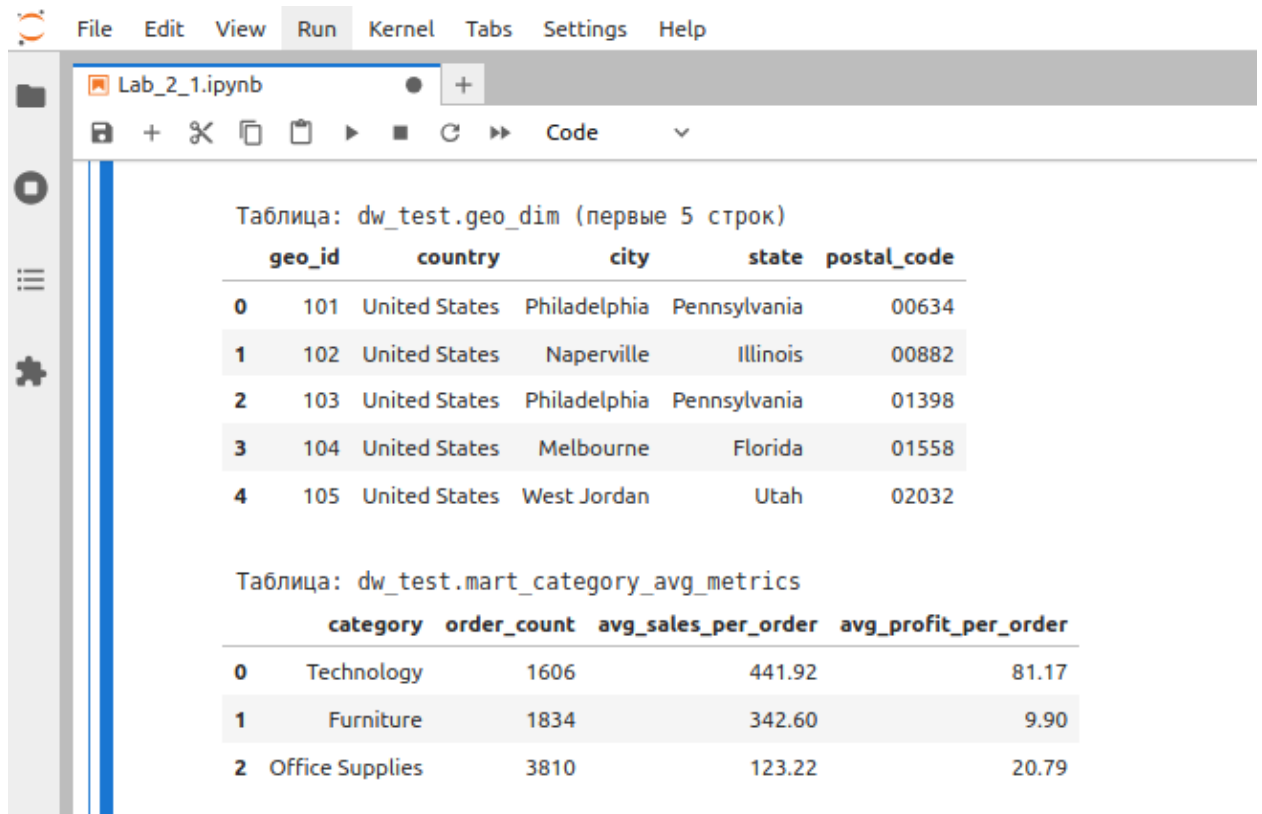


Таблица: dw_test.geo_dim (первые 5 строк)

	geo_id	country	city	state	postal_code
0	101	United States	Philadelphia	Pennsylvania	00634
1	102	United States	Naperville	Illinois	00882
2	103	United States	Philadelphia	Pennsylvania	01398
3	104	United States	Melbourne	Florida	01558
4	105	United States	West Jordan	Utah	02032

Таблица: dw_test.mart_category_avg_metrics

	category	order_count	avg_sales_per_order	avg_profit_per_order
0	Technology	1606	441.92	81.17
1	Furniture	1834	342.60	9.90
2	Office Supplies	3810	123.22	20.79

4. Вывод

dbt упрощает построение DWH, так как позволяет описывать трансформации данных декларативно на SQL с использованием версионизируемых моделей, а не вручную управлять DDL/DML скриптами. DBT обеспечивает автоматическое управление зависимостями между таблицами, документирование и тестирование данных, их визуализация в lineage. В итоге он кардинально повышает надежность, поддерживаемость и доверие к данным.