

ДЕПАРТАМЕНТ ОБРАЗОВАНИЯ И НАУКИ ГОРОДА МОСКВЫ

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

(ГАОУ ВО МГПУ)

Институт цифрового образования

Департамент информатики, управления и технологий

Лабораторная работа № 2.1

Вариант 30

по дисциплине «Платформы Data Engineering»

Выполнил:

студент группы БД-251м

Направление подготовки/Специальность

38.04.05 - Бизнес-информатика

Трухачев Никита Алексеевич

(Ф.И.О.)

Проверил:

Доцент, к.т.н

(ученая степень, звание)

Босенко Тимур Муртазович

(Ф.И.О.)

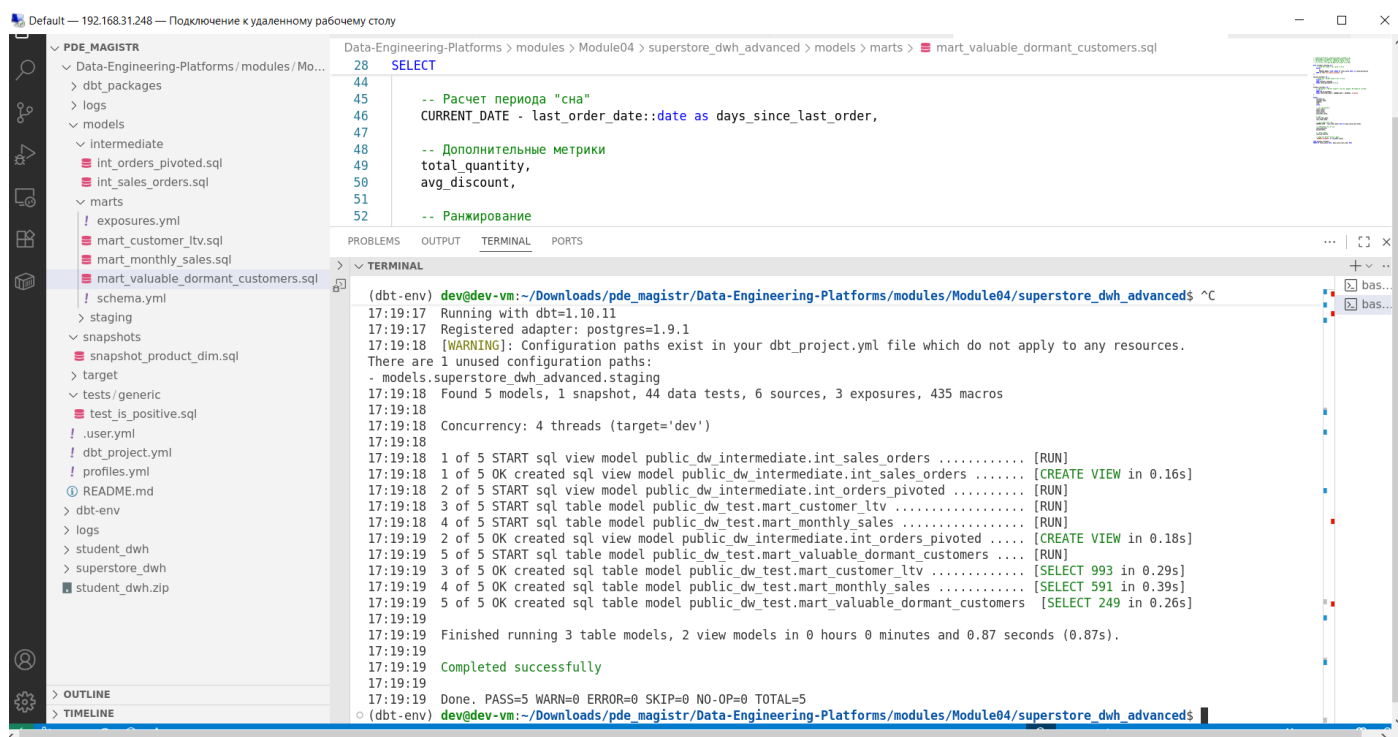
Москва 2025

Краткое описание архитектуры проекта

Слой `intermediate` служит для подготовки и преобразования данных после стадии `staging`, но до финальных витрин. Здесь выполняются сложные джойны, агрегации и бизнес-преобразования, которые являются общими для нескольких витрин. Например, в `int_orders_pivoted.sql` данные о заказах, трансформируются в удобную для анализа структуру. Это позволяет избежать дублирования логики и повышает переиспользуемость кода.

Слой `marts` содержит готовые к использованию аналитические витрины, ориентированные на конкретные бизнес-сценарии, такие как анализ продаж (`mart_monthly_sales`) или расчёт LTV клиентов (`mart_customer_ltv`). Разделение на `intermediate` и `marts` необходимо для соблюдения принципа единой ответственности: промежуточный слой отвечает за техническую подготовку данных, а витрины – за предоставление понятных и целостных бизнес-метрик конечным потребителям (аналитикам, отчетам).

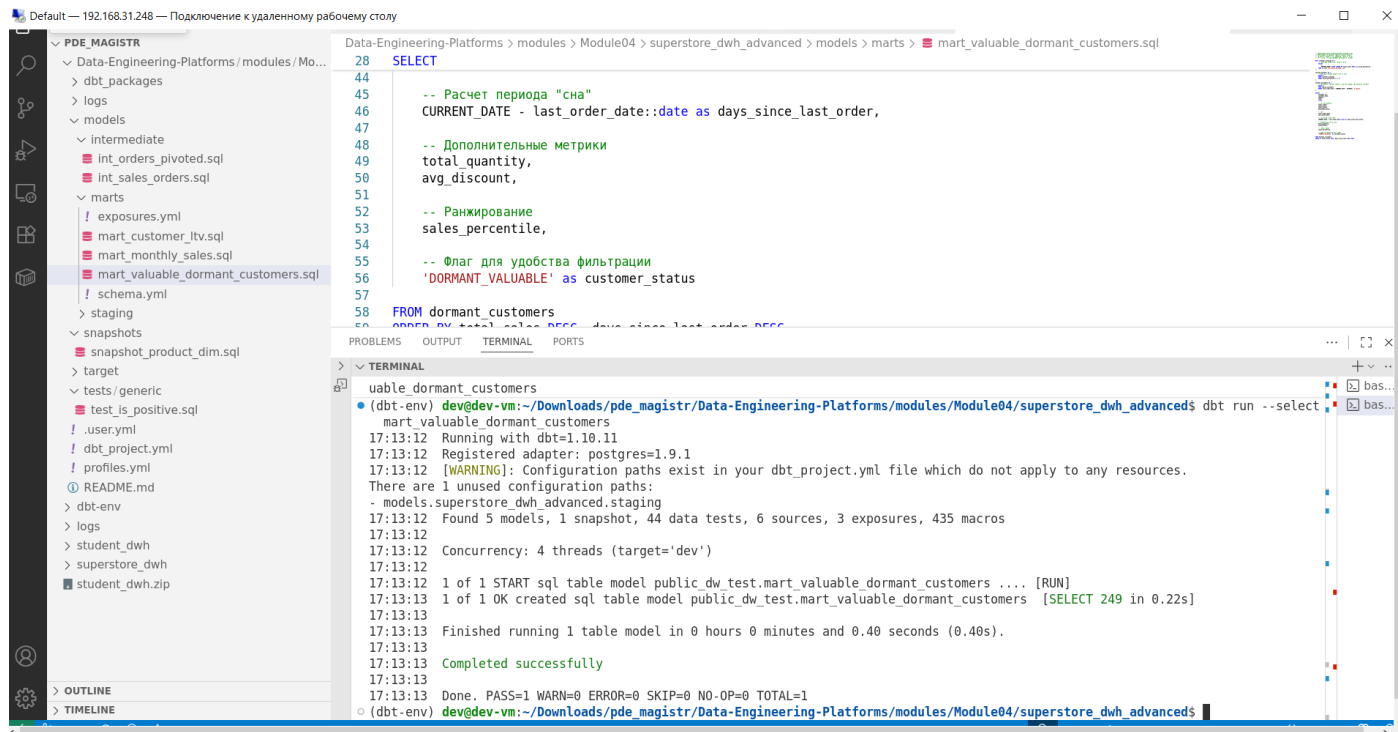
dbt run



```
28 SELECT
44
45 -- Расчет периода "сна"
46 CURRENT_DATE - last_order_date::date as days_since_last_order,
47
48 -- Дополнительные метрики
49 total_quantity,
50 avg_discount,
51
52 -- Ранжирование
```

```
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$ ^C
17:19:17 Running with dbt=1.10.11
17:19:17 Registered adapter: postgres=1.9.1
17:19:18 [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.superstore_dwh_advanced.staging
17:19:18 Found 5 models, 1 snapshot, 44 data tests, 6 sources, 3 exposures, 435 macros
17:19:18
17:19:18 Concurrency: 4 threads (target='dev')
17:19:18
17:19:18 1 of 5 START sql view model public_dw_intermediate.int_sales_orders ..... [RUN]
17:19:18 1 of 5 OK created sql view model public_dw_intermediate.int_sales_orders ..... [CREATE VIEW in 0.16s]
17:19:18 2 of 5 START sql view model public_dw_intermediate.int_orders_pivoted ..... [RUN]
17:19:18 3 of 5 START sql table model public_dw_test.mart_customer_ltv ..... [RUN]
17:19:18 4 of 5 START sql table model public_dw_test.mart_monthly_sales ..... [RUN]
17:19:18 2 of 5 OK created sql view model public_dw_intermediate.int_orders_pivoted ..... [CREATE VIEW in 0.18s]
17:19:18 5 of 5 START sql table model public_dw_test.mart_valuable_dormant_customers .... [RUN]
17:19:18 3 of 5 OK created sql table model public_dw_test.mart_customer_ltv ..... [SELECT 993 in 0.29s]
17:19:18 4 of 5 OK created sql table model public_dw_test.mart_monthly_sales ..... [SELECT 591 in 0.39s]
17:19:18 5 of 5 OK created sql table model public_dw_test.mart_valuable_dormant_customers [SELECT 249 in 0.26s]
17:19:19
17:19:19 Finished running 3 table models, 2 view models in 0 hours 0 minutes and 0.87 seconds (0.87s).
17:19:19
17:19:19 Completed successfully
17:19:19
17:19:19 Done. PASS=5 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=5
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$
```

dbt run --select mart_valuable_dormant_customers



```
28 SELECT
44
45 -- Расчет периода "сна"
46 CURRENT_DATE - last_order_date::date as days_since_last_order,
47
48 -- Дополнительные метрики
49 total_quantity,
50 avg_discount,
51
52 -- Ранжирование
53 sales_percentile,
54
55 -- Флаг для удобства фильтрации
56 'DORMANT VALUABLE' as customer_status
57
58 FROM dormant_customers
59 ORDER BY total_sales DESC, days_since_last_order DESC
```

```
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$ dbt run --select
mart_valuable_dormant_customers
17:13:12 Running with dbt=1.10.11
17:13:12 Registered adapter: postgres=1.9.1
17:13:12 [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.superstore_dwh_advanced.staging
17:13:12 Found 5 models, 1 snapshot, 44 data tests, 6 sources, 3 exposures, 435 macros
17:13:12
17:13:12 Concurrency: 4 threads (target='dev')
17:13:12
17:13:12 1 of 1 START sql table model public_dw_test.mart_valuable_dormant_customers .... [RUN]
17:13:13 1 of 1 OK created sql table model public_dw_test.mart_valuable_dormant_customers [SELECT 249 in 0.22s]
17:13:13
17:13:13 Finished running 1 table model in 0 hours 0 minutes and 0.40 seconds (0.40s).
17:13:13
17:13:13 Completed successfully
17:13:13
17:13:13 Done. PASS=1 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=1
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$
```

dbt test

```
Default — 192.168.31.248 — Подключение к удаленному рабочему столу
Data-Engineering-Platforms > modules > Module04 > superstore_dwh_advanced > models > marts > mart_valuable_dormant_customers.sql
28 SELECT
44
45
PROBLEMS OUTPUT TERMINAL PORTS
> TERMINAL
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$ dbt test
17:19:53 28 of 44 PASS not_null_mart_valuable_dormant_customers_avg_order_value ..... [PASS in 0.22s]
17:19:53 33 of 44 START test_not_null_mart_valuable_dormant_customers_days_since_last_order [RUN]
17:19:53 34 of 44 START test_not_null_mart_valuable_dormant_customers_first_order_date .. [RUN]
17:19:53 31 of 44 PASS not_null_mart_valuable_dormant_customers_customer_name ..... [PASS in 0.13s]
17:19:53 35 of 44 START test_not_null_mart_valuable_dormant_customers_last_order_date ... [RUN]
17:19:53 32 of 44 PASS not_null_mart_valuable_dormant_customers_customer_status ..... [PASS in 0.12s]
17:19:53 36 of 44 START test_not_null_mart_valuable_dormant_customers_sales_percentile .. [RUN]
17:19:53 33 of 44 PASS not_null_mart_valuable_dormant_customers_days_since_last_order ... [PASS in 0.15s]
17:19:53 34 of 44 PASS not_null_mart_valuable_dormant_customers_first_order_date ..... [PASS in 0.15s]
17:19:53 37 of 44 START test_not_null_mart_valuable_dormant_customers_segment ..... [RUN]
17:19:53 38 of 44 PASS not_null_mart_valuable_dormant_customers_state ..... [PASS in 0.14s]
17:19:53 35 of 44 PASS not_null_mart_valuable_dormant_customers_last_order_date ..... [PASS in 0.14s]
17:19:53 39 of 44 START test_not_null_mart_valuable_dormant_customers_total_orders ..... [RUN]
17:19:53 36 of 44 PASS not_null_mart_valuable_dormant_customers_sales_percentile ..... [PASS in 0.15s]
17:19:53 40 of 44 START test_not_null_mart_valuable_dormant_customers_total_profit ..... [RUN]
17:19:53 37 of 44 PASS not_null_mart_valuable_dormant_customers_segment ..... [PASS in 0.14s]
17:19:53 39 of 44 PASS not_null_mart_valuable_dormant_customers_total_orders ..... [PASS in 0.13s]
17:19:53 38 of 44 PASS not_null_mart_valuable_dormant_customers_state ..... [PASS in 0.17s]
17:19:53 41 of 44 START test_not_null_mart_valuable_dormant_customers_total_quantity .... [RUN]
17:19:53 42 of 44 START test_not_null_mart_valuable_dormant_customers_total_sales ..... [RUN]
17:19:53 43 of 44 START test_unique_mart_customer_ltv_customer_id ..... [RUN]
17:19:53 40 of 44 PASS not_null_mart_valuable_dormant_customers_total_profit ..... [PASS in 0.17s]
17:19:53 44 of 44 START test_mart_valuable_dormant_customers_customer_id ..... [RUN]
17:19:53 41 of 44 PASS not_null_mart_valuable_dormant_customers_total_sales ..... [PASS in 0.16s]
17:19:53 42 of 44 PASS not_null_mart_valuable_dormant_customers_total_quantity ..... [PASS in 0.17s]
17:19:53 43 of 44 PASS unique_mart_customer_ltv_customer_id ..... [PASS in 0.17s]
17:19:53 44 of 44 PASS unique_mart_valuable_dormant_customers_customer_id ..... [PASS in 0.08s]
17:19:53
17:19:53 Finished running 44 data tests in 0 hours 0 minutes and 2.52 seconds (2.52s).
17:19:53
17:19:53 Completed successfully
17:19:53
17:19:53 Done. PASS=44 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=44
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$
```

dbt test --select mart_pareto_customer_analysis

```
Default — 192.168.31.248 — Подключение к удаленному рабочему столу
Data-Engineering-Platforms > modules > Module04 > superstore_dwh_advanced > models > marts > mart_valuable_dormant_customers.sql
28 SELECT
44
45
PROBLEMS OUTPUT TERMINAL PORTS
> TERMINAL
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$ dbt test --select mart_valuable_dormant_customers
17:20:51 9 of 23 PASS not_null_mart_valuable_dormant_customers_city ..... [PASS in 0.17s]
17:20:51 13 of 23 START test_not_null_mart_valuable_dormant_customers_days_since_last_order [RUN]
17:20:51 11 of 23 PASS not_null_mart_valuable_dormant_customers_customer_name ..... [PASS in 0.12s]
17:20:51 14 of 23 START test_not_null_mart_valuable_dormant_customers_first_order_date ... [RUN]
17:20:51 12 of 23 PASS not_null_mart_valuable_dormant_customers_customer_status ..... [PASS in 0.12s]
17:20:51 10 of 23 PASS not_null_mart_valuable_dormant_customers_customer_id ..... [PASS in 0.16s]
17:20:51 15 of 23 START test_not_null_mart_valuable_dormant_customers_last_order_date ... [RUN]
17:20:51 16 of 23 START test_not_null_mart_valuable_dormant_customers_sales_percentile .. [RUN]
17:20:51 13 of 23 PASS not_null_mart_valuable_dormant_customers_days_since_last_order ... [PASS in 0.19s]
17:20:51 17 of 23 START test_not_null_mart_valuable_dormant_customers_segment ..... [RUN]
17:20:51 14 of 23 PASS not_null_mart_valuable_dormant_customers_first_order_date ..... [PASS in 0.15s]
17:20:51 18 of 23 START test_not_null_mart_valuable_dormant_customers_state ..... [RUN]
17:20:51 15 of 23 PASS not_null_mart_valuable_dormant_customers_last_order_date ..... [PASS in 0.12s]
17:20:51 16 of 23 PASS not_null_mart_valuable_dormant_customers_sales_percentile ..... [PASS in 0.14s]
17:20:51 19 of 23 START test_not_null_mart_valuable_dormant_customers_total_orders ..... [RUN]
17:20:51 20 of 23 START test_not_null_mart_valuable_dormant_customers_total_profit ..... [RUN]
17:20:51 17 of 23 PASS not_null_mart_valuable_dormant_customers_segment ..... [PASS in 0.11s]
17:20:51 21 of 23 START test_not_null_mart_valuable_dormant_customers_total_quantity .... [RUN]
17:20:51 18 of 23 PASS not_null_mart_valuable_dormant_customers_state ..... [PASS in 0.12s]
17:20:51 22 of 23 START test_not_null_mart_valuable_dormant_customers_total_sales ..... [RUN]
17:20:51 20 of 23 PASS not_null_mart_valuable_dormant_customers_total_profit ..... [PASS in 0.13s]
17:20:51 19 of 23 PASS not_null_mart_valuable_dormant_customers_total_orders ..... [PASS in 0.17s]
17:20:51 23 of 23 START test_unique_mart_valuable_dormant_customers_customer_id ..... [RUN]
17:20:51 21 of 23 PASS not_null_mart_valuable_dormant_customers_total_quantity ..... [PASS in 0.13s]
17:20:51 22 of 23 PASS not_null_mart_valuable_dormant_customers_total_sales ..... [PASS in 0.10s]
17:20:51 23 of 23 PASS unique_mart_valuable_dormant_customers_customer_id ..... [PASS in 0.06s]
17:20:52
17:20:52 Finished running 23 data tests in 0 hours 0 minutes and 1.24 seconds (1.24s).
17:20:52
17:20:52 Completed successfully
17:20:52
17:20:52 Done. PASS=23 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=23
```

dbt snapshot

The screenshot shows a VS Code editor with a project structure on the left and a terminal window on the right. The project structure includes folders like 'logs', 'models', 'intermediate', 'marts', 'staging', 'snapshots', 'target', and 'tests/generic'. The terminal window shows the output of the 'dbt snapshot' command, which includes a warning about unused configuration paths and a successful completion message.

```
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$ dbt docs serve
return func(*args, **kwargs)
File "/home/dev/Downloads/pde_magistr/dbt-env/lib/python3.10/site-packages/dbt/cli/requires.py", line 303, in wrapper
return func(*args, **kwargs)
File "/home/dev/Downloads/pde_magistr/dbt-env/lib/python3.10/site-packages/dbt/cli/requires.py", line 350, in wrapper
return func(*args, **kwargs)
File "/home/dev/Downloads/pde_magistr/dbt-env/lib/python3.10/site-packages/dbt/cli/main.py", line 307, in docs_serve
results = task.run()
File "/home/dev/Downloads/pde_magistr/dbt-env/lib/python3.10/site-packages/dbt/task/docs/serve.py", line 29, in run
httpd.serve_forever()
File "/usr/lib/python3.10/socketserver.py", line 232, in serve_forever
ready = selector.select(poll_interval)
File "/usr/lib/python3.10/selectors.py", line 416, in select
fd_event_list = self._selector.poll(timeout)
KeyboardInterrupt

• (dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$ dbt snapshot
17:27:38 Running with dbt=1.10.11
17:27:38 Registered adapter: postgres=1.9.1
17:27:38 [WARNING]: Configuration paths exist in your dbt_project.yml file which do not apply to any resources.
There are 1 unused configuration paths:
- models.superstore_dwh_advanced.staging
17:27:39 Found 5 models, 1 snapshot, 44 data tests, 6 sources, 3 exposures, 435 macros
17:27:39
17:27:39 Concurrency: 4 threads (target='dev')
17:27:39
17:27:39 1 of 1 START snapshot dw_snapshots.snapshot_product_dim ..... [RUN]
17:27:39 1 of 1 OK snapshot dw_snapshots.snapshot_product_dim ..... [SELECT 4544 in 0.16s]
17:27:39
17:27:39 Finished running 1 snapshot in 0 hours 0 minutes and 0.37 seconds (0.37s).
17:27:39
17:27:39 Completed successfully
17:27:39
17:27:39 Done. PASS=1 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=1
(dbt-env) dev@dev-vm:~/Downloads/pde_magistr/Data-Engineering-Platforms/modules/Module04/superstore_dwh_advanced$
```

Архитектура DWH

Lineage graph



Скриншот с данными из индивидуальной mart-модели. Запрос

```
try:
    print("\n Спящие ценные клиенты. Определяет клиентов из топ-25% по общей выручке, которые не совершали покупок последние 6 месяцев")
    df_facts = pd.read_sql("SELECT * FROM public_dw_test.mart_valuable_dormant_customers;", engine)
    display(df_facts)
except Exception as e:
    print(f"❌ Не удалось загрузить dw_test.sales_fact: {e}")
```

Скриншот с данными из индивидуальной mart-модели. Ответ

Спящие ценные клиенты. Определяет клиентов из топ-25% по общей выручке, которые не совершали покупок последние 6 месяцев																
	customer_id	customer_name	segment	city	state	total_orders	total_sales	total_profit	avg_order_value	first_order_date	last_order_date	days_since_last_order	total_quantity	avg_discount	sales_percentile	customer_status
0	TC-20980	Tamara Chand	Consumer	Decatur	Alabama	5	56368.0100	26617.4351	1761.500313	2016-11-07	2018-11-26	2520	123	0.106250	0.000000	DORMANT_VALUABLE
1	RB-19360	Raymond Buch	Consumer	Auburn	California	6	44764.3600	20798.2696	1065.818095	2018-04-01	2019-09-25	2217	164	0.100000	0.001008	DORMANT_VALUABLE
2	TA-21385	Tom Ashbrook	Consumer	Chicago	Illinois	4	43729.4600	14100.4589	1507.912414	2016-09-12	2019-10-22	2190	103	0.082759	0.002016	DORMANT_VALUABLE
3	SC-20095	Sanjit Chand	Consumer	Concord	Arkansas	9	39919.1680	16417.5512	725.803055	2016-02-12	2019-01-15	2470	209	0.061818	0.003024	DORMANT_VALUABLE
4	AB-10105	Adrian Barton	Consumer	Bloomington	Arizona	10	39124.4880	15770.2795	954.255805	2016-12-20	2019-11-19	2162	168	0.253659	0.004032	DORMANT_VALUABLE
...
244	NS-18640	Noel Staavos	Consumer	Baltimore	California	13	8513.4010	-348.0493	113.512013	2016-06-25	2019-07-09	2295	326	0.200000	0.245968	DORMANT_VALUABLE
245	SC-20020	Sam Craven	Consumer	Houston	Michigan	5	8508.4412	-265.8690	181.030664	2016-03-03	2017-11-10	2901	172	0.192340	0.246976	DORMANT_VALUABLE
246	ND-18460	Neil Ducich	Consumer	Chandler	Alabama	6	8470.3750	1256.4941	403.351190	2016-02-06	2019-06-30	2304	93	0.100000	0.247984	DORMANT_VALUABLE
247	AP-10915	Arthur Pritchep	Consumer	Columbus	California	10	8463.4400	1551.5263	111.361053	2016-08-23	2019-09-23	2219	257	0.086842	0.248992	DORMANT_VALUABLE
248	MN-17935	Michael Nguyen	Consumer	Clinton	Maryland	6	8457.5580	1411.3529	352.398250	2016-02-16	2019-11-26	2155	90	0.250000	0.250000	DORMANT_VALUABLE
249 rows x 16 columns																

249 rows x 16 columns

Вывод. В чем преимущество использования промежуточных моделей и витрин по сравнению с работой напрямую с единой таблицей фактов?

Использование промежуточных моделей и витрин данных предоставляет ключевое преимущество в виде декомпозиции сложности. Вместо работы с единой громоздкой таблицей фактов, которая содержит всю сырую информацию, сложные бизнес-преобразования разбиваются на управляемые этапы. Промежуточные модели (intermediate) абстрагируют техническую сложность – такие операции, как джойны нескольких таблиц, очистка данных или предварительные агрегации – что делает логику прозрачнее и значительно упрощает тестирование и повторное использование кода.