HURTOWNIE DANYCH

Projekt – Analiza danych platformy e-commerce Olist (Brazylia)

ALEKSANDER STEPANIUK
NR. INDEKSU: 272644
Politechnika Wrocławska, Informatyka Stosowana

Etap 2 – 26.05.2025 r.

1. SC_CreateStageTables

Tworzy schemat Stage oraz wszystkie tymczasowe tabele, do których będą wczytywane surowe pliki CSV i później przechowywane dane oczyszczone:

- CreateStageSchema
  - jeżeli nie istnieje tworzy schema Stage
- CreateStageOrders
  - tworzy Stage.Orders z olist_orders.csv
- CreateStageOrderItems
  - tworzy Stage.OrderItems z olist_order_items.csv
- CreateStagePayments
  - tworzy Stage.Payments z olist_order_payments.csv
- CreateStageReviews
  - tworzy Stage.Reviews z olist_order_reviews.csv
- CreateStageCustomers
  - tworzy Stage.Customers z olist_customers.csv
- CreateStageSellers
  - tworzy Stage.Sellers z olist_sellers.csv
- CreateStageProducts
  - tworzy Stage.Products z olist_products.csv
- CreateStageProductCategoryNameTranslated
  - tworzy Stage.ProductCategoryNameTranslation z tłumaczeniem nazw kategorii
- CreateStageCities
  - tworzy Stage.Cities z danymi brazylijskich miast
- CreateStageOrdersClean
  - tworzy tabelę Stage.OrdersClean na wyniki oczyszczania dat i miar.
- CreateStageCustomersClean
  - tworzy Stage.CustomersClean na wzbogacone dane klientów o dane miast
- CreateStageSellersClean
  - tworzy Stage.SellersClean na wzbogacone dane sprzedawców o dane miast
- CreateStageProductsClean
  - tworzy Stage.ProductsClean na wzbogacone dane produktów

2. SC_LoadStageData

Wczytuje dane z plików CSV do tabel Stage.[…] za pomocą zadań Data Flow:

- DFT_LoadOrdersStage
- DFT_LoadOrderItemsStage
- DFT_LoadPaymentsStage
- DFT_LoadReviewsStage
- DFT_LoadCustomersStage
- DFT_LoadSellersStage
- DFT_LoadProductsStage
- DFT_LoadCitiesStage
- DFT_LoadProductCategoryNameTranslationStage

3. SC_CleanStage
Oczyszcza i wzbogaca dane:

- DFT_OrdersClean
    - konwersja dat na DATETIME, obliczenie czasu dostawy - delivery_time (w dniach), zapis do Stage.OrdersClean
- DFT_CustomersClean
    - fuzzy lookup miast, dodanie populacji, powierzchni, gęstości zaludnienia, zapis do Stage.CustomersClean
- DFT_SellersClean
    - analogiczne wzbogacenie danych sprzedawców, zapis do Stage.SellersClean
- DFT_ProductsClean
    - tłumaczenie kategorii, konwersja zmiennych, zapis do Stage.ProductsClean

4. SC_CreateFinalSchemaTables
Tworzy schemat o nazwie „Stepaniuk” oraz wszystkie tabele docelowe hurtowni:

- CreateSchema
    - tworzy schemat Stepaniuk jeśli nie istnieje.
- CreateMonthDim
    - Tworzy Stepaniuk.MonthDim z 12 wierszami
- CreateWeekdayDim
    - Tworzy Stepaniuk.WeekdayDim z 7 wierszami
- CreateTimeDim
    - Tworzy Stepaniuk.TimeDim (kluczem czas + dodatkowe atrybuty do dni tygodnia, miesiąca, roku itd.)
- CreateCustomerDim
    - Tworzy Stepaniuk.CustomerDim (klient + dane miast)
- CreateSellerDim
    - Tworzy Stepaniuk.SellerDim (sprzedawca + dane miast)
- CreateProductDim
    - Tworzy Stepaniuk.ProductDim
- CreatePaymentDim
    - Tworzy Stepaniuk.PaymentDim (unikalne metody płatności)
- CreateReviewDim
    - Tworzy Stepaniuk.ReviewDim
- CreateFactOrders
    - Tworzy Stepaniuk.FactOrders (na poziomie pojedynczych produktów order_item)

5. SC_LoadFinalData

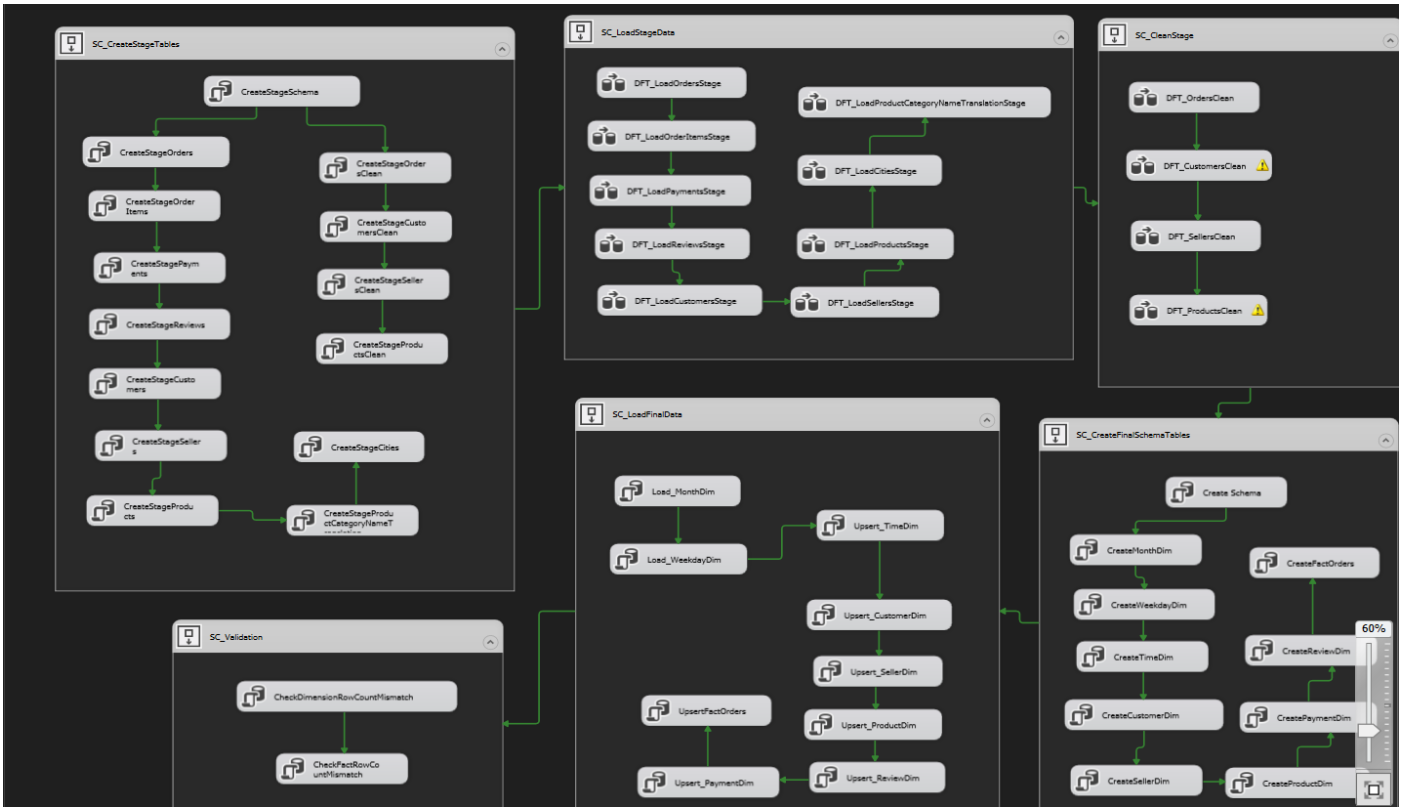
Ładuje dane do wymiarów i faktów, stosując przyrostowe upserty (insert where not exists) przy pomocy MERGE:

- LoadMonthDim
- LoadWeekdayDim
- UpsertTimeDim
- UpsertCustomerDim
- UpsertSellerDim
- UpsertProductDim
- UpsertReviewDim
- UpsertPaymentDim
- UpsertFactOrders

6. SC_Validation

Weryfikuje poprawność załadowanych danych i w razie błędów Raisuje błąd.

- CheckDimensionRowCountMismatch
- CheckFactRowCountMismatch



Lp.	Źródłowy plik	Źródłowa kolumna	Docelowa kolumna	Typ danych
1	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.full_datetime	DATETIME
2	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.time_key	BIGINT
3	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.year_n	SMALLINT
4	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.quarter_n	SMALLINT
5	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.month_key	SMALLINT
6	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.day_n	SMALLINT
7	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.weekday_key	SMALLINT
8	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.hour_n	SMALLINT
9	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.minute_n	SMALLINT
10	olist_orders_dataset.csv	order_purchase_timestamp	TimeDim.second_n	SMALLINT
11	olist_customers_dataset.csv	customer_id	CustomerDim.customer_id	VARCHAR(50)
12	olist_customers_dataset.csv	customer_state	CustomerDim.customer_state	CHAR(2)
13	olist_customers_dataset.csv	customer_city	CustomerDim.customer_city	VARCHAR(100)
14	brazilian_cities.csv	IBGE_RES_POP	CustomerDim.city_population	INT
15	brazilian_cities.csv	AREA	CustomerDim.city_area_km2	DECIMAL(10,2)
16	brazilian_cities.csv	(computed) population/area	CustomerDim.city_density	DECIMAL(10,2)
17	olist_sellers_dataset.csv	seller_id	SellerDim.seller_id	VARCHAR(50)
18	olist_sellers_dataset.csv	seller_state	SellerDim.seller_state	CHAR(2)
19	olist_sellers_dataset.csv	seller_city	SellerDim.seller_city	VARCHAR(100)
20	brazilian_cities.csv	IBGE_RES_POP	SellerDim.city_population	INT
21	brazilian_cities.csv	AREA	SellerDim.city_area_km2	DECIMAL(10,2)
22	brazilian_cities.csv	(computed) population/area	SellerDim.city_density	DECIMAL(10,2)
23	olist_products_dataset.csv	product_id	ProductDim.product_id	VARCHAR(50)
24	olist_products_dataset.csv	product_category_name + translation	ProductDim.category	VARCHAR(100)
25	olist_products_dataset.csv	(z CSV tłumaczeń) product_category_name_english	ProductDim.category	VARCHAR(100)
26	olist_products_dataset.csv	product_category_name (podkategoria)	ProductDim.sub_category	VARCHAR(100)
27	olist_order_payments_dataset.csv	payment_type	PaymentDim.payment_type	VARCHAR(50)
28	olist_order_payments_dataset.csv	payment_type	PaymentDim.payment_type_key	INT (surrogate)
29	olist_order_reviews_dataset.csv	review_id	ReviewDim.review_id	VARCHAR(50)
30	olist_order_reviews_dataset.csv	review_score	ReviewDim.review_score	SMALLINT

31	olist_order_reviews_dataset.csv	review_comment_message	ReviewDim.review_comment	TEXT
32	olist_order_reviews_dataset.csv	review_creation_date	ReviewDim.review_date	DATE
33	olist_order_items_dataset.csv	order_item_id	FactOrders.order_item_id	VARCHAR(50)
34	olist_order_items_dataset.csv	order_id	FactOrders.order_id	VARCHAR(50)
35	olist_orders_clean (Stage.OrdersClean)	order_purchase_timestamp	FactOrders.average_delivery_time	DECIMAL(10,2)
36	olist_order_items_dataset.csv + freight	price + freight_value	FactOrders.total_revenue	DECIMAL(18,2)
37	olist_order_items_dataset.csv	order_item_id	FactOrders.total_items	INT
38	olist_order_reviews_dataset.csv	review_score	FactOrders.average_review_score	DECIMAL(3,2)
39	olist_order_payments_dataset.csv (p.seq=1)	payment_sequential	FactOrders.payment_sequential	SMALLINT
40	olist_order_payments_dataset.csv (p.seq=1)	payment_installments	FactOrders.payment_installments	SMALLINT
41	olist_order_payments_dataset.csv (p.seq=1)	payment_value	FactOrders.payment_value	DECIMAL(18,2)
42	Stage.OrdersClean	time_key	FactOrders.time_key	BIGINT
43	Stage.OrdersClean	customer_id	FactOrders.customer_id	VARCHAR(50)
44	Stage.OrderItems	seller_id	FactOrders.seller_id	VARCHAR(50)
45	Stage.OrderItems	product_id	FactOrders.product_id	VARCHAR(50)
46	Stepaniuk.PaymentDim	payment_type_key	FactOrders.payment_type_key	INT