

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

A screenshot of a computer

AI-generated content may be incorrect.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |