HURTOWNIE DANYCH

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

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

Etap I – 19.05.2025 r.

1. **Tytuł projektu**

   Analiza danych platformy e-commerce Olist (Brazylia)

2. **Charakterystyka dziedziny problemowej, krótki opis obszaru analizy, problemy i potrzeby**

   Celem projektu jest analiza brazylijskiego rynku e-commerce, gdzie platforma Olist pośredniczy między tysiącami sprzedawców, a setkami tysięcy klientów, gromadząc dane o zamówieniach, płatnościach, produktach i opiniach.

   W obszarze analizy ważne jest śledzenie przebiegu transakcji od momentu złożenia zamówienia aż po dostawę i ocenę satysfakcji użytkownika. Rozproszone w różnych plikach typu CSV źródła danych utrudniają szybkie agregowanie takich informacji. Problemy pojawiają się także przy zapewnieniu spójnej jakości danych – niekompletne lub niespójne rekordy recenzji czy płatności mogą zafałszować raporty, a brak zintegrowanego modelu danych wydłuża czas przygotowania analiz.

   Potrzeby biznesowe koncentrują się na możliwości błyskawicznego generowania wielowymiarowych raportów (np. przychód wg regionu i miesiąca, ocena sprzedawcy według kwartału), monitorowaniu najważniejszych wskaźników jakości obsługi klienta (czas dostawy, liczba reklamacji) oraz elastycznej segmentacji klientów i produktów, co wymaga wdrożenia hurtowni danych z jasno zdefiniowanymi wymiarami i faktami.

3. **Cel przedsięwzięcia (oczekiwania) oraz zakres analizy – badane aspekty**

   Cel główny: zbudowanie hurtowni danych, która pozwoli na:

   - Monitorowanie kluczowych wskaźników sprzedaży (przychód, liczba zamówień) w czasie.
   - Analizę jakości dostaw i satysfakcji klienta (recenzje, czasy realizacji).
   - Podział klientów i sprzedawców wg regionów i zachowań zakupowych.

   Zakres:

   - Dane sprzedażowe, zamówienia, płatności, opinie, dane klientów, sprzedawców, produktów.
   - Analizy czasowe (miesiąc, kwartał), geograficzne (stan, miasto), produktowe (kategorie).

4. **Źródła danych (lokalizacja, format, dostępność), wstępna analiza źródeł danych**

   Dane zostały pobrane ze strony kaggle.com:

   - https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Lp.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1	olist_customers_dataset	.csv	99441	9.03	zamówienia (daty: złożenia, zatwierdzenia, dostawy)
2	olist_geolocation _dataset	.csv	1000164	61.27	pozycje zamówień (produkt, ilość, cena)
3	olist_order_items _dataset	.csv	112650	15.44	klienci (id, miasto, stan, kod pocztowy)
4	olist_order_payments _dataset	.csv	103887	5.78	płatności (metoda, rata, wartość)
5	olist_order_reviews_dataset	.csv	103887	14.45	opinie klientów (ocena, komentarz, data)
6	olist_orders_dataset	.csv	99442	17.65	produkty (id, kategoria, wymiary)
7	olist_products	.csv	32952	2.38	sprzedawcy (id, lokalizacja)
8	olist_sellers	.csv	3096	0.17	tłumaczenia kategorii produktowych
9	product_category_name_translation	.csv	72	0.26	geolokalizacja kodów pocztowych

5. Profilowanie danych (analiza jakości danych oraz ich przydatności w projekcie)

Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1	**order_purchase_timestamp**	datetime	2016-09-04 – 2018-09-03	brak nulli, format ISO spójny w 100 % wierszy
2	**order_approved_at**	datetime	2016-09-04 – 2018-09-05	~0,1 % null (zamówienia anulowane)
3	**order_delivered_carrier_date**	datetime	2016-09-07 – 2018-09-17	~0,2 % null (problemy logistyczne)
4	**order_delivered_customer_date**	datetime	2016-09-09 – 2018-09-23	~0,3 % null (zwrócone lub nie dostarczone)
5	**order_status**	string	delivered, shipped, invoiced, created, approved…	brak nulli; wartości spójne
6	**price**	decimal(10,2)	0.01 – 9999.00	brak wartości ujemnych, ~0,01 % skrajnie niskich cen (promocje)
7	**payment_type**	string	credit_card, boleto, voucher, debit_card	~5 % null (zwroty/refundy); pozostałe wartości zgodne z dokumentacją
8	**payment_installments**	integer	1 – 12	brak nulli, realistyczny rozkład (najwięcej 1–3 raty)
9	**review_score**	integer	1 – 5	~0,3 % null (brak opinii), średnia ocena ≈ 4,09
10	**review_creation_date**	datetime	2016-10-01 – 2018-10-15	~0,3 % null, daty recenzji mieszczą się do 30 dni po dostawie
11	**customer_state**	string (2)	SP, RJ, MG, BA, CE, …	brak nulli, 27 kodów stanów (BR-XX), wszystkie poprawne zgodnie z ISO 3166-2:BR
12	**customer_city**	string	São Paulo, Rio de Janeiro, Salvador, …	~0,05 % literówek (akcenty), można ujednolicić wielkość liter
13	**seller_state**	string (2)	SP, RJ, MG, PR, RS, …	brak nulli; ~3 095 unikalnych sprzedawców, wszystkie stany pokryte

14	**seller_city**	string	São Paulo, Curitiba, Porto Alegre, …	ok, podobnie jak w kliencie: drobne literówki/różnice w zapisie
15	**product_category_name**	string	bed_bath_table, health_beauty, sports_leisure, …	71 kategorii, wszystkie występują min. raz, brak nulli
16	**product_weight_g**	integer	50 – 40000	~0,1 % null, wartości realistyczne, możliwe outliery do weryfikacji
17	**product_length_cm**	integer	5 – 200	~0,1 % null, typowe zakresy dla e-commerce
18	**product_height_cm**	integer	1 – 150	analogicznie do długości
19	**product_width_cm**	integer	2 – 150	ok, można obliczyć objętość
20	**geolocation_lat**	decimal(9,6)	-33.868820 – 5.193082	~2 % błędnych koordynat (poza granicami BR) – wymaga filtrowania
21	**geolocation_lng**	decimal(9,6)	-73.985506 – -34.793129	jak wyżej

6. Definicja typów encji/klas (wraz z właściwościami) oraz związków pomiędzy nimi, diagram klas (propozycja wymiarów, hierarchii, miar addytywnych i nieaddytywnych)

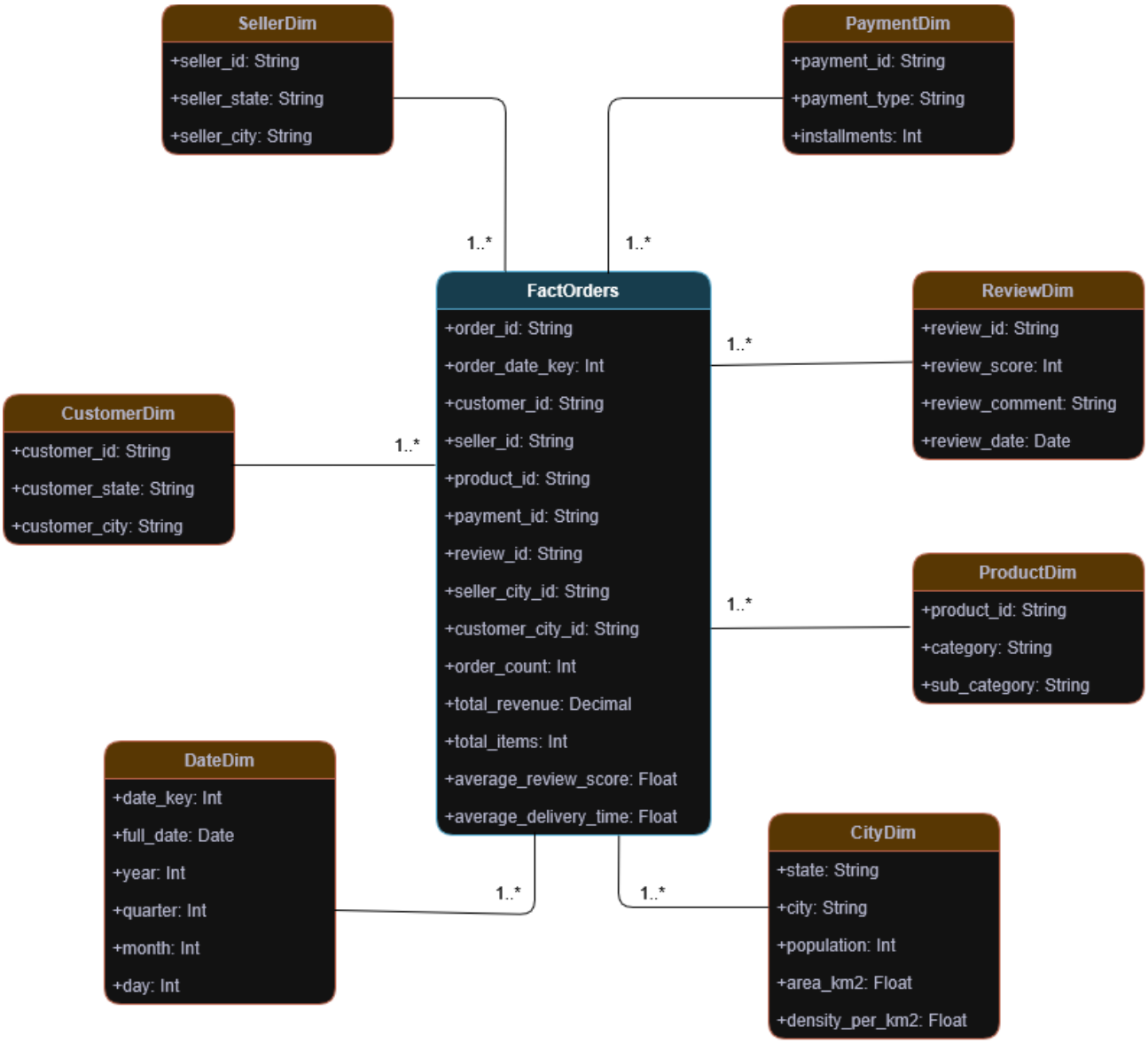
Encje wymiarów:

1. DateDim
    - Klucz: date_key (INT, YYYYMMDD)
    - Atrybuty: full_date, year, quarter, month, day, weekday
    - Hierarchia: Year -> Quarter -> Month -> Day -> Weekday
2. CustomerDim
    - Klucz: customer_id (VARCHAR)
    - Atrybuty: customer_city_key, customer_zip_code, customer_segment
    - Hierarchia: State -> City -> Zip Code -> Segment
3. SellerDim
    - Klucz: seller_id (VARCHAR)
    - Atrybuty: seller_city_key, seller_zip_code, seller_segment
    - Hierarchia: State -> City -> Zip Code -> Segment
4. ProductDim
    - Klucz: product_id (VARCHAR)
    - Atrybuty: category, sub_category, weight_g, height_cm, width_cm, length_cm
    - Hierarchia: Category -> Sub-category -> Product
5. PaymentDim
    - Klucz: payment_id (VARCHAR)
    - Atrybuty: payment_type, installments, payment_value
    - Hierarchia: Type -> Installments
6. ReviewDim
    - Klucz: review_id (VARCHAR)
    - Atrybuty: review_score, review_date, review_comment
    - Hierarchia: Score -> Date
7. CityDim
    - Klucz: city_key (INT, autoincrement)
    - Atrybuty: state, city, population, area_km2, density_per_km2
    - Hierarchia: State -> City

Encja faktu:

7. FactOrders
    - Klucz główny: order_id (VARCHAR)
    - Klucze obce:
        - order_date_key -> DateDim
        - customer_id -> CustomerDim
        - seller_id -> SellerDim
        - product_id -> ProductDim
        - payment_id -> PaymentDim
        - review_id -> ReviewDim
        - customer_city_key -> CityDim
        - seller_city_key -> CityDim
    - Miary addytywne:
        - order_count (INT) – liczba zamówień,
        - total_revenue (DECIMAL) – suma przychodu,
        - total_items (INT) – liczba produktów.
    - Miary nieaddytywne:
        - average_review_score (DECIMAL) – średnia ocena;
        - payment_type (VARCHAR), order_status (VARCHAR) – opisowe, nie sumują się.

Diagram klas:



7. Min. 10 wielowymiarowych zestawień, które zostaną utworzone po wdrożeniu kostki

1. Przychód i liczba zamówień według miesiąca i stanu klienta
2. Średnia ocena i liczba opinii wg sprzedawcy i kwartału
3. Rozkład typów płatności wg kategorii produktu i roku
4. Średni czas dostawy wg regionu sprzedawcy i miesiąca
5. Top 10 produktów wg przychodu i liczby sztuk w pewnym analizowanym okresie
6. Średni czas dostawy według gęstości zaludnienia
7. Liczba zamówień według populacji miasta
8. Liczba zamówień nowych vs powracających klientów według miast
9. Przychód według dnia tygodnia i typu płatności
10. Top 10 najgorszych sprzedawców według średniej ocen i miesiąca

8. Implementacja bazy danych zgodnie z zaproponowanym konceptualnym modelem danych

```sql
-- schemat
CREATE SCHEMA Stepaniuk;

-- tabela pomocnicza dla miesięcy
CREATE TABLE Stepaniuk.MonthDim (
    month_key    SMALLINT    PRIMARY KEY,
    month_name   VARCHAR(20) NOT NULL
);

INSERT INTO Stepaniuk.MonthDim VALUES
(1,'January'),  (2,'February'), (3,'March'),    (4,'April'),
(5,'May'),      (6,'June'),     (7,'July'),     (8,'August'),
(9,'September'),(10,'October'), (11,'November'),(12,'December');

-- wymiar daty
CREATE TABLE Stepaniuk.DateDim (
    date_key      INT          PRIMARY KEY,
    full_date     DATE         NOT NULL,
    year_n        SMALLINT     NOT NULL,
    quarter_n     SMALLINT     NOT NULL,
    month_key     SMALLINT     NOT NULL       REFERENCES Stepaniuk.MonthDim(month_key),
    day_n         SMALLINT     NOT NULL
);

-- wymiar klienta
CREATE TABLE Stepaniuk.CustomerDim (
    customer_id   VARCHAR(50)  PRIMARY KEY,
    customer_state CHAR(2)     NOT NULL,
    customer_city  VARCHAR(100) NOT NULL
);

-- wymiar sprzedawcy
CREATE TABLE Stepaniuk.SellerDim (
    seller_id     VARCHAR(50)  PRIMARY KEY,
    seller_state  CHAR(2)      NOT NULL,
    seller_city   VARCHAR(100) NOT NULL
);
```

```sql
-- wymiar produktu
CREATE TABLE Stepaniuk.ProductDim (
    product_id     VARCHAR(50)  PRIMARY KEY,
    category       VARCHAR(100) NOT NULL,
    sub_category   VARCHAR(100)
);

-- wymiar płatności
CREATE TABLE Stepaniuk.PaymentDim (
    payment_id     VARCHAR(50)  PRIMARY KEY,
    payment_type   VARCHAR(30)  NOT NULL,
    installments   SMALLINT     NOT NULL
);

-- wymiar opinii
CREATE TABLE Stepaniuk.ReviewDim (
    review_id      VARCHAR(50) PRIMARY KEY,
    review_score   SMALLINT    NOT NULL,
    review_comment TEXT,
    review_date    DATE        NOT NULL
);

-- tabela faktów
CREATE TABLE Stepaniuk.FactOrders (
    order_id              VARCHAR(50)  PRIMARY KEY,
    order_date_key        INT          NOT NULL          REFERENCES Stepaniuk.DateDim(date_key),
    customer_id           VARCHAR(50)  NOT NULL          REFERENCES Stepaniuk.CustomerDim(customer_id),
    seller_id             VARCHAR(50)  NOT NULL          REFERENCES Stepaniuk.SellerDim(seller_id),
    product_id            VARCHAR(50)  NOT NULL          REFERENCES Stepaniuk.ProductDim(product_id),
    payment_id            VARCHAR(50)  NOT NULL          REFERENCES Stepaniuk.PaymentDim(payment_id),
    review_id             VARCHAR(50)                    REFERENCES Stepaniuk.ReviewDim(review_id),
    order_count           INT          NOT NULL DEFAULT 1,
    total_revenue         DECIMAL(18,2)NOT NULL,
    total_items           INT          NOT NULL,
    average_review_score  DECIMAL(3,2),
    average_delivery_time DECIMAL(10,2) -- dni
);
```

9. Wnioski

Zdecydowana większość atrybutów jest kompletna i może z powodzeniem trafić do hurtowni danych – mamy pełne informacje o zamówieniach, klientach, produktach i płatnościach, co pozwala na zbudowanie rozbudowanych wymiarów czasowego, geograficznego, klienta, sprzedawcy, produktu i płatności. Dane o opiniach są niemal kompletne, choć kilkaset rekordów nie zawiera ocen lub komentarzy, co jednak nie powinno zaburzyć ogólnych trendów. Z kolei geolokalizacje wymagają odfiltrowania kilku procent współrzędnych spoza terytorium Brazylii, ale same kody pocztowe umożliwiają precyzyjne grupowanie według stanów i miast.

Wartości numeryczne – takie jak cena, liczba rat czy wymiary produktów – mieszczą się w sensownych zakresach i nie zawierają błędnych skrajnych wartości, co czyni je gotowymi do agregacji i obliczeń KPI. Potencjalne outliery w wadze lub wymiarach można prawdopodobnie wyfiltrować. Pola tekstowe (kategorie, nazwy miast) wymagają jedynie podstawowej normalizacji (usunięcie literówek, standaryzacja akcentów), by zapobiec duplikacji wymiaru.

W wymiarach znajdziemy naprawdę sporo do badania: czasowego (analiza sezonowości i trendów), przestrzennego (różnice między regionami), produktowego (popularność i marże w kategoriach) oraz behawioralnego (liczba rat czy typ płatności jako wskaźniki preferencji klientów). Faktowe miary – przychód, liczba zamówień, średnia ocena czy czas dostawy – pozwolą na wielowymiarowe zestawienia i dogłębne analizy jakości obsługi. Dzięki temu hurtownia stanie się solidnym fundamentem dla raportów sprzedażowych, monitoringu satysfakcji klientów oraz optymalizacji procesów logistycznych i marketingowych.