МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Практикум №2

з курсу «Програмування інтелектуальних інформаційних систем»

на тему: «Створення і наповнення сховища даних»

Викладач:
Курченко О.А.

Виконав:
Хільчук А.В.
студент 3 курсу
групи ІП-14 ФІОТ

Київ-2023

Практична робота №1

Тема: Створення і наповнення сховища даних
Завдання:

1. Відповідно до заданого варіанту підготувати необхідні дані у виглядітаблиць MS Excel і зберегти їх як персональні файли. Для підготовки даних використовувати тематичні сайти Інтернет, результати проходження практик, довідники і каталоги.
2. Створити персональне сховище даних і організувати доступ до нього.
3. Виконати завантаження даних з таблиць MS Excel за допомогою Майстраімпорту.
4. Організувати завантаження даних за допомогою Майстра експорту в сховище даних.
5. Здійснити наступний імпорт даних з сховища: кількість відвантаженого товару в розрізі дат та товарів по вибраному Вами клієнтові, залишивши одну властивість товару (вибір властивості довільний).

Галузь – торгівля побутовою технікою.

Виконання:

1. Відповідно до заданого варіанту підготувати необхідні дані у вигляді таблиць MS Excel і зберегти їх як персональні файли. Для підготовки даних використовувати тематичні сайти Інтернет, результати проходження практик, довідники і каталоги.

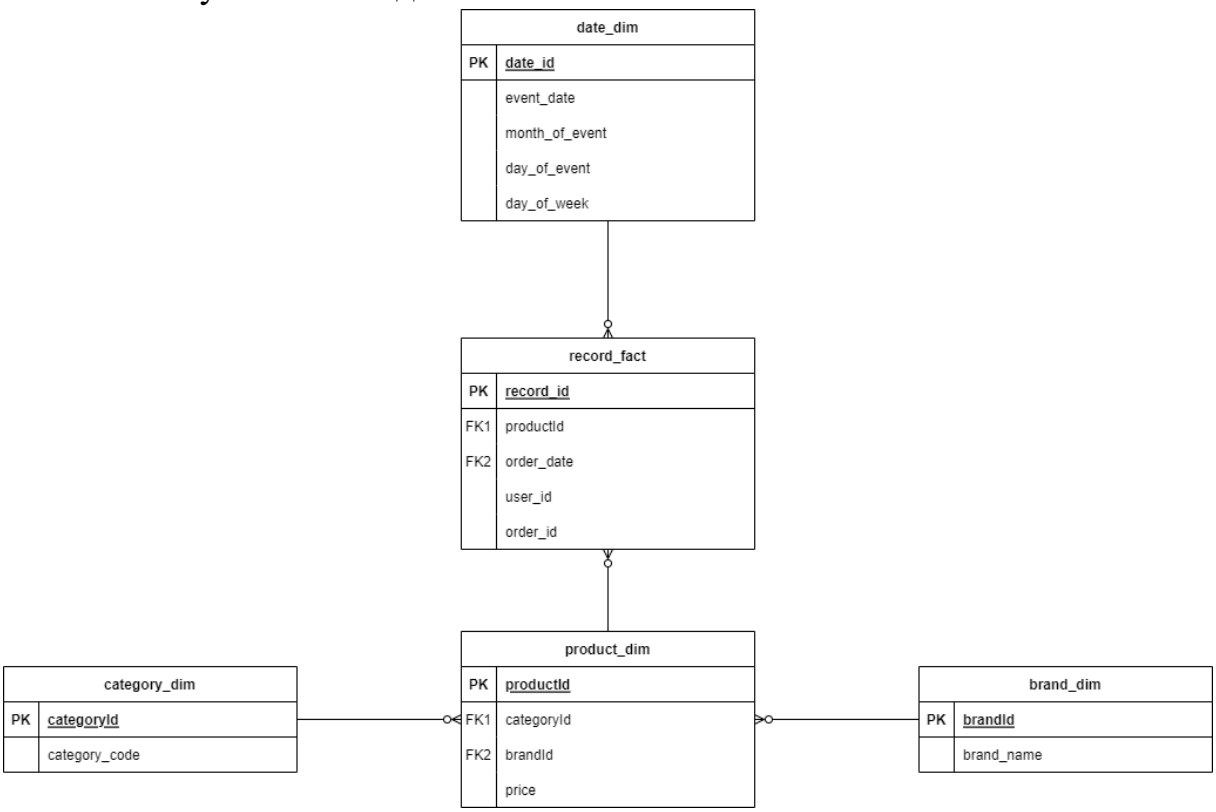
   Для виконання даної лабораторної роботи було взято даний набір даних:
   https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-electronics-store

   Даний датасет містить інформацію про усі записи замовльен в онлайн мережі для покупок електронних твораів. Однак не всі записи відповідають покупкам побутової техніки, отож, попередньо за допомогою скрипту, наведеного в **додатку А,** було відібрано записи, що

задовольняють предметну область. Також у процесі було переформатовано стовпець дати та відкинуто рядки з відсутніми значеннями.

2. Створити персональне сховище даних та організувати доступ до нього.

Результатом проектування є сховище даних за типом “сніжинка” матиме наступний вигляд:



category_dim – таблиця виміру категорії товару. Містить код та назву категорії
date_dim – таблиця з датами
brand_dim – таблиця виміру бренду товару. Містить код та назву бренду
product_dim – таблиця виміру товару. Містить код товару, його вартість, а також посилання на категорію та бренд
record_fact – фактова таблиця замовлення. Містить код замовлення, дату замовлення, ідентифікатор замовлення, ідентифікатор покупця, а також посилання на товар.

Реалізовано сховище буде за допомогою СУБД MS SQL Server. Скрипт для створення таблиць наведено в **додатку Б**

3. Виконати завантаження даних з таблиць MS Excel за допомогою Майстра імпорту

З міркувань зручності та безпеки, спочатку дані з .csv файлу буде імпортовану в стейдж зону даного вигляду

stage_zone
product_id
order_id
category_id
event_time
category_code
brand
price
user_id

Скрипт для її створення наведено в **додатку В.**

Тоді за допомогою майстра імпорту було проведено заповнення даними.

	event_time	order_id	product_id	category_id	category_code	brand	price	user_id
1	2020-04-26	1	2872	32	appliances.kitchen.refrigerators	lg	462,94	7422
2	2020-04-26	2	915	42	appliances.personal.scales	polaris	30,07	5265
3	2020-04-29	3	1451	49	appliances.kitchen.kettle	tefal	7,85	5514
4	2020-04-29	4	605	54	appliances.personal.scales	polaris	21,97	926
5	2020-04-29	5	2916	58	appliances.kitchen.blender	polaris	43,96	5024
6	2020-04-29	6	421	42	appliances.personal.scales	scarlett	18,5	3943
7	2020-04-29	7	1547	33	appliances.kitchen.refrigerators	caso	312,48	5536
8	2020-04-29	8	727	18	appliances.iron	tefal	81	8429
9	2020-04-29	9	267	56	appliances.kitchen.mixer	maxwell	16,18	6642
10	2020-04-29	10	160	61	appliances.kitchen.kettle	vitek	20,81	8460
11	2020-04-29	11	135	57	appliances.kitchen.meat_grinder	moulinex	57,85	3412
12	2020-04-29	12	2412	43	appliances.personal.hair_cutter	imetec	20,81	8238
13	2020-04-29	13	1532	29	appliances.environment.air_heater	ava	11,55	8054
14	2020-04-29	14	102	13	appliances.kitchen.washer	samsung	451,37	759
15	2020-04-29	14	1182	57	appliances.kitchen.meat_grinder	philips	115,72	759
16	2020-04-29	14	561	50	appliances.kitchen.microwave	samsung	122,66	759
17	2020-04-29	15	2319	58	appliances.kitchen.blender	polaris	32,38	8410
18	2020-04-29	16	87	50	appliances.kitchen.microwave	ava	43,96	8133
19	2020-04-29	17	2366	56	appliances.kitchen.mixer	polaris	46,27	3333
20	2020-04-29	18	2858	61	appliances.kitchen.kettle	philips	32,38	2328
21	2020-04-29	19	225	25	appliances.environment.vacuum	thomas	15,02	8483
22	2020-04-29	20	327	27	appliances.environment.water_heater	ariston	118,03	8480
23	2020-04-29	21	51	43	appliances.personal.hair_cutter	philips	46,27	973
24	2020-04-29	22	315	22	appliances.environment.vacuum	samsung	69,42	6590
25	2020-04-29	23	799	58	appliances.kitchen.blender	maxwell	23,13	8495
26	2020-04-29	24	40	43	appliances.personal.hair_cutter	philips	85,63	2031

4. Організувати завантаження даних за допомогою Майстра експорту в сховище даних.

Перш за все заповнюємо brand_dim та category_dim. brand_dim заселяємо унікальними значеннями:

	brandId	brand_name
1	1	wmf
2	2	kerasys
3	3	compliment
4	4	elica
5	5	bellissima
6	6	bosch
7	7	whirlpool
8	8	samsung
9	9	huter
10	10	delonghi
11	11	atmor
12	12	hansa
13	13	fissman
14	14	birjusa
15	15	aoki
16	16	scarlett
17	17	thermex

Однак під час аналізу даних категорій та їх кодів виникає цікава ситуація:

```
unique values in category_code: 30
unique values in category_id: 77
```

Найменувань категорій більше ніж їх кодів. На жаль, ніяких уточнень стосовно цього не було знайдено, однак можна припустити, що різні за ідентифікатором категорії є просто підтипами основних категорій.
Однак з даними припущеннями робити нічого, а при аналізі даних масове дублювання значень може повторюватися, було прийнято рішення розділяти категорії суто за їх кодами.
Отож, таблицю category_dim буде заповнено унікальними значеннями category_code:

	categoryId	category_code
1	1	appliances.kitchen.oven
2	2	appliances.steam_cleaner
3	3	appliances.kitchen.blender
4	4	appliances.kitchen.mixer
5	5	appliances.kitchen.fryer
6	6	appliances.kitchen.meat_grinder
7	7	appliances.kitchen.toster
8	8	appliances.kitchen.steam_cooker
9	9	appliances.kitchen.grill
10	10	appliances.kitchen.refrigerators
11	11	appliances.kitchen.kettle
12	12	appliances.environment.air_heater
13	13	appliances.kitchen.juicer
14	14	appliances.kitchen.coffee_machine
15	15	appliances.personal.scales
16	16	appliances.personal.hair_cutter
17	17	appliances.kitchen.hood
18	18	appliances.environment.vacuum
19	19	appliances.ironing_board
20	20	appliances.environment.fan
21	21	appliances.kitchen.dishwasher
22	22	appliances.environment.climate
23	23	appliances.personal.massager
24	24	appliances.sewing_machine
25	25	appliances.environment.water_heater
26	26	appliances.iron
27	27	appliances.kitchen.coffee_grinder
28	28	appliances.environment.air_conditioner
29	29	appliances.kitchen.microwave
30	30	appliances.kitchen.washer

Тоді заповнюємо product_dim:

	productId	categoryId	brandId	price
1	1	16	64	71.74
2	2	18	79	611.09
3	3	18	79	425.90
4	4	15	32	18.50
5	5	26	79	115.72
6	6	16	16	12.71
7	7	9	79	231.46
8	8	30	8	613.40
9	9	16	89	15.02
10	10	25	115	122.66
11	11	28	119	138.87
12	12	26	79	32.38
13	13	18	6	101.83
14	14	21	12	300.90
15	15	10	119	162.01
16	16	25	11	48.59
17	17	25	29	162.01
18	18	24	33	175.90
19	19	30	46	439.79
20	20	25	29	85.63
21	21	1	6	462.94
22	22	25	17	99.51
23	23	25	116	101.83

Опісля займаємося date_dim:

	date_id	event_date	month_of_event	day_of_event	day_of_week
1	1	2020-01-05	1	5	Sunday
2	2	2020-01-06	1	6	Monday
3	3	2020-01-07	1	7	Tuesday
4	4	2020-01-08	1	8	Wednesday
5	5	2020-01-09	1	9	Thursday
6	6	2020-01-10	1	10	Friday
7	7	2020-01-11	1	11	Saturday
8	8	2020-01-12	1	12	Sunday
9	9	2020-01-13	1	13	Monday
10	10	2020-01-14	1	14	Tuesday
11	11	2020-01-15	1	15	Wednesday
12	12	2020-01-16	1	16	Thursday
13	13	2020-01-17	1	17	Friday
14	14	2020-01-18	1	18	Saturday
15	15	2020-01-19	1	19	Sunday
16	16	2020-01-20	1	20	Monday
17	17	2020-01-21	1	21	Tuesday
18	18	2020-01-22	1	22	Wednesday
19	19	2020-01-23	1	23	Thursday
20	20	2020-01-24	1	24	Friday
21	21	2020-01-25	1	25	Saturday
22	22	2020-01-26	1	26	Sunday
23	23	2020-01-27	1	27	Monday
24	24	2020-01-28	1	28	Tuesday
25	25	2020-01-29	1	29	Wednesday
26	26	2020-01-30	1	30	Thursday
27	27	2020-01-31	1	31	Friday
28	28	2020-02-01	2	1	Saturday
29	29	2020-02-02	2	2	Sunday
30	30	2020-02-03	2	3	Monday
31	31	2020-02-04	2	4	Tuesday
32	32	2020-02-05	2	5	Wednesday
33	33	2020-02-06	2	6	Thursday
34	34	2020-02-07	2	7	Friday
35	35	2020-02-08	2	8	Saturday
36	36	2020-02-09	2	9	Sunday

I record_fact:

	record_id	order_id	productId	order_date	user_id
1	1	1	2872	2020-04-26	7422
2	2	2	915	2020-04-26	5265
3	3	3	1451	2020-04-29	5514
4	4	4	605	2020-04-29	926
5	5	5	2916	2020-04-29	5024
6	6	6	421	2020-04-29	3943
7	7	7	1547	2020-04-29	5536
8	8	8	727	2020-04-29	8429
9	9	9	267	2020-04-29	6642
10	10	10	160	2020-04-29	8460
11	11	11	135	2020-04-29	3412
12	12	12	2412	2020-04-29	8238
13	13	13	1532	2020-04-29	8054
14	14	14	102	2020-04-29	759
15	15	14	1182	2020-04-29	759
16	16	14	561	2020-04-29	759
17	17	15	2319	2020-04-29	8410
18	18	16	87	2020-04-29	8133
19	19	17	2366	2020-04-29	3333
20	20	18	2858	2020-04-29	2328
21	21	19	225	2020-04-29	8483
22	22	20	327	2020-04-29	8480
23	23	21	51	2020-04-29	973
24	24	22	315	2020-04-29	6590
25	25	23	799	2020-04-29	8495
26	26	24	40	2020-04-29	2031
27	27	25	2918	2020-04-29	8482

Скрипт для заповнення таблиць наведений у **додатку Г**.

5. Здійснити наступний імпорт даних з сховища: кількість відвантаженого товару в розрізі дат та товарів по вибраному Вами клієнтові, залишивши одну властивість товару (вибір властивості довільний).

   Для виконання даного завдання було застосовано платформу інтелектуального аналізу даних Power BI.

productId	12 квітня 2020 р.	13 квітня 2020 р.	18 травня 2020 р.	19 червня 2020 р.	17 липня 2020 р.	**Total**
11	1					**1**
24				1		**1**
211		1				**1**
238				1		**1**
265		1				**1**
683					1	**1**
908	1					**1**
1247				1		**1**
1250				1		**1**
2355			1			**1**
Total	**2**	**2**	**1**	**4**	**1**	**10**

Дана матриця демонструє кількості куплених товарів по датах для покупця з ідентифікатором 25323, з вартістю товарів більшою за 120

Висновок

Отож, у ході виконання лабораторної роботи було взято набір даних про покупки електронних товарів у онлайн магазині та підготовлено їх для подальшого аналізу і зберігання. Дані імпортовано та завантажено до спроектованого сховища даних типу "сніжинка" за допомогою СУБД MS SQL Server. Результуюче сховище даних включає таблиці для зберігання інформації про категорії товарів, бренди, конкретні товари, дати та факти замовлень. Під час заповнення таблиць було виявлено особливості у деяких наявних даних, зокрема категорій, що зрезультувало у розділення їх за їх кодами для уникнення масового дублювання значень. Урешті-решт із використанням платформи інтелектуального аналізу даних Power BI було побудувано матрицю кількостей куплених товарів по датах для конкретного клієнта з певною вартістю товарів. Набуто практичних навичок проектування сховищ даних та побудови BI ріщень.

Додаток А

Скрипт для фільтрації вхідних даних

```
import pandas as pd

if __name__=="__main__":
    df=pd.read_csv("data.csv")
    df=df.dropna()

    df["event_time"]=df["event_time"].str.split(" ").str[0]

    selected=df[df['category_code'].str.contains('appliances')]

    slected=df[df["event_time"]>"2019-12-31"]

    int_columns = ['order_id', 'product_id', 'category_id', 'user_id']
    for col in int_columns:
        selected[col]=selected.groupby(col).ngroup() + 1
    selected.to_csv('result_data.csv',index=False, encoding='utf-
8',line_terminator='\r\n')
```

Додаток Б

Скрипт для створення таблиць сховища даних

```
--warehouse
CREATE TABLE date_dim (
       date_id int PRIMARY KEY IDENTITY(1,1),
    event_date DATE,
    month_of_event INT,
    day_of_event INT,
    day_of_week VARCHAR(20)
);
CREATE TABLE category_dim (
    categoryId INT PRIMARY KEY IDENTITY(1,1),
    category_code VARCHAR(255) NOT NULL
);

CREATE TABLE brand_dim (
    brandId INT PRIMARY KEY IDENTITY(1,1),
    brand_name VARCHAR(255) NOT NULL
);

CREATE TABLE product_dim (
    productId INT PRIMARY KEY,
    categoryId INT NOT NULL,
    brandId INT NOT NULL,
    price DECIMAL(10, 2) NOT NULL,
    FOREIGN KEY (categoryId) REFERENCES category_dim(categoryId),
    FOREIGN KEY (brandId) REFERENCES brand_dim(brandId)
);


CREATE TABLE record_fact (
    record_id INT PRIMARY KEY IDENTITY(1,1),
       order_id INT NOT NULL,
    productId INT NOT NULL,
    order_date DATE NOT NULL,
    user_id INT NOT NULL,
    FOREIGN KEY (productId) REFERENCES product_dim(productId),
);
```

Додаток В

Скрипт для створення таблиці стейдж зони

```
-- stage zone
CREATE TABLE stage (
    event_time DATE,
    order_id INT,
    product_id INT,
    category_id INT,
    category_code VARCHAR(255),
    brand VARCHAR(255),
    price DECIMAL(18, 2),
    user_id INT
);
```

Додаток Г

Скрипт для заповнення таблиць сховища даних

```
insert into category_dim
select distinct [category_code] from stage
select * from category_dim

insert into brand_dim
select distinct brand from stage
select * from brand_dim

INSERT INTO product_dim
SELECT DISTINCT s.product_id, c.categoryId, b.brandId, s.price
FROM stage s
JOIN category_dim c ON s.category_code = c.category_code
JOIN brand_dim b ON s.brand = b.brand_name order by product_id

INSERT INTO date_dim
SELECT DISTINCT
    event_time,
    MONTH(event_time),
    DAY(event_time),
    DATENAME(weekday, event_time)
FROM stage order by event_time;

select * from date_dim

INSERT INTO record_fact
SELECT s.order_id, s.product_id, date_dim.date_id, s.user_id
FROM stage s
JOIN date_dim ON s.event_time = date_dim.event_date order by event_date;
select * from record_fact
```