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 №1

з курсу «Аналіз даних в інформаційнних системах» на тему: «Створення сховища даних»

Викладач:

Олійник Ю.О.

Виконав:

студент 2 курсу групи ІП-15 ФІОТ  
Мєшков Андрій Ігорович

Київ-2023

**Практикум №1**

**Створення сховища даних**

**Мета роботи**: ознайомитись з підходами до створення сховищ даних.

**Завдання:** Навчитися створювати процедури завантаження даних до сховища.

1. Самостійно обрати не менше 3-х джерел відкритих даних.
2. Спроектувати модель Stage зони для ETL процесів.
3. Спроектувати модель основного сховища за типом «зірка» або «сніжинка».
4. Створити ETL засоби:
   * завантажити дані до Stage зони
   * створити набір процедур/функцій для перетворення та завантаження даних до основного сховища (або створити засобами програмних ETL засобів). Передбачити можливість завантаження змінених та додаткових даних.
5. Завантажити дані до основного сховища даних.

**Хід роботи:**

1. Для виконання лабораторної роботи було обрано 3 джерела відкритих даних на сайті <https://www.kaggle.com//>. А саме:

* Звіт про щастя у світі: https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report.csv
* Зміна клімату: дані про температуру поверхні Землі: https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalLandTemperaturesByCountry.csv
* Тероризм: https://www.kaggle.com/datasets/START-UMD/gtd

Предметною областю лабораторної роботи є рівень щастя людей та залежність щастя від навколишніх змін.

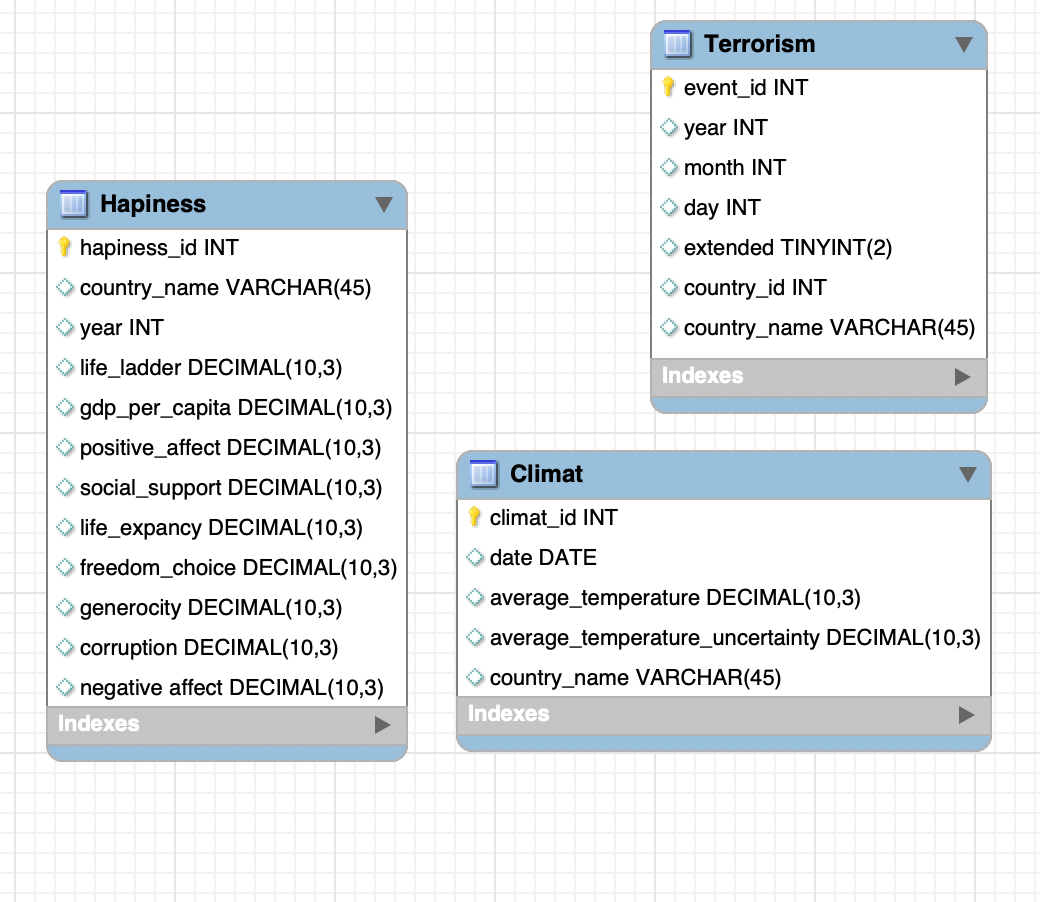
Нижче наведені поля для кожного з файлів, які безпосередньо використовувались у подальшій побудові бізнес-процесів:

*Таблиця 1 - поля вхідних файлів*

|  |  |  |
| --- | --- | --- |
| world-happiness-report.csv | country\_name | Назва країни |
| year | Рік |
| life\_ladder | Життєва драбина |
| gdp\_per\_capita | ВВП на душу населення |
| social\_support | Соціальна допомога |
| life\_expancy | Очікувана здорова тривалість життя при народженні |
| freedom\_choice | Свобода робити життєвий вибір |
| generocity | Щедрість |
| corruption | Сприйняття корупції |
| positive\_affect | Позитивний ефект |
| negative affect | Негативний ефект |
| GlobalLandTemperaturesBy-Country.csv | date | Дата |
| average\_temperature | Середня температура |
| average\_temperature\_uncertainty | Невизначеність середньої температури |
| country\_name | Назва країни |
| globalterrorismdb\_0718-dist.csv | event\_id | Номер події |
| year | Рік |
| month | Місяць |
| day | День |
| extended | Розширення, тривання події більш ніж 24 години |
| country\_id | Номер країни |
| country\_name | Назва країни |

1. Модель Stage зони для ETL процесів

В результаті розробки була спроектована схема stage-зони, яка зображена на рисунку 2.1. Дана модель відображає таблиці для даних із вхідних джерел.



*Рисунок 2.1 – Stage зона для ETL процесів*

Скрипти створення Stage зони знаходяться у додатку А.

Опис таблиць stage зони:

1. Hapiness – призначена для зберігання інформації про щастя країн в різні роки.
2. Terrorism – призначена для зберігання інформації про терористичні акти в різних країнах в різні роки.
3. Climat – призначена для зберігання інформації про зміну температури в країнах.
4. Модель основного сховища за типом «зірка» або «сніжинка»

У процесі розробки моделі сховища даних було створено одну таблицю фактів та 4 таблиці вимірів:

-dim\_climat – таблиця виміру клімату за температурою;

-dim\_terrorism – таблиця виміру терористичних атак;

-dim\_date – таблиця виміру дати;

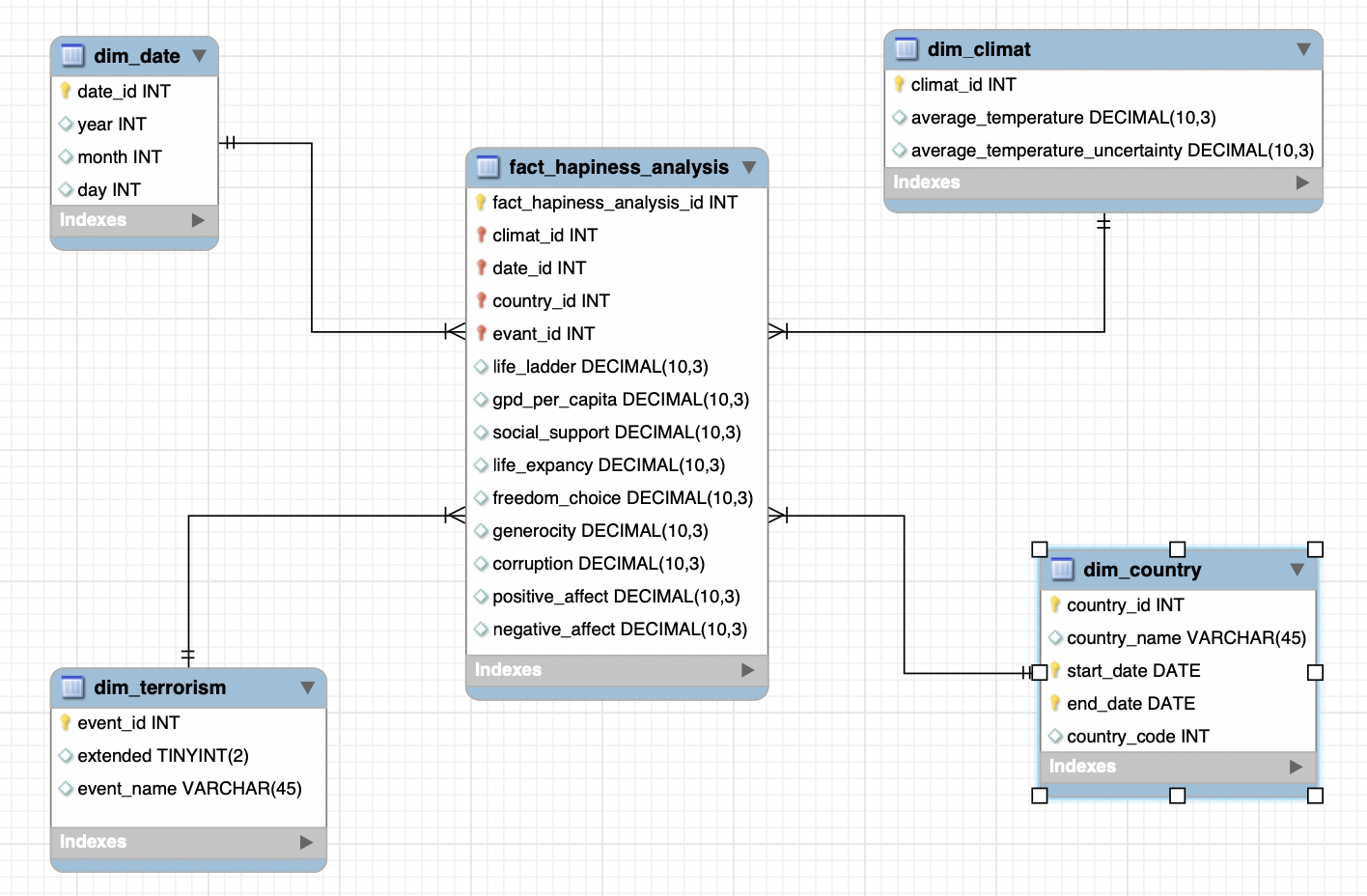
-dim\_country – таблиця виміру країни;

-fact\_hapiness\_analysis – таблиця фактів щастя в точці часу та простору;

*Таблиця 3.1 – Таблиця атрибутів таблиць сховища даних*

|  |  |  |  |
| --- | --- | --- | --- |
| **Назва таблиці** | **Назва атрибуту** | **Тип даних** | **Первинний ключ** |
| dim\_climat | climat\_id | INT | climat\_id |
| average\_temperature | DECIMAL(10,3) |
| average\_temperature\_ uncertainty | DECIMAL(10,3) |
| dim\_terrorism | event\_id | INT | event\_id |
| event\_name | VARCHAR(45) |
| extended | TINYINT(2) |
| dim\_date | date\_id | INT | date\_id |
| year | INT |
| month | INT |
| day | INT |
| dim\_country | country\_id | INT | country\_id |
| country\_code | INT |
| country\_name | VARCHAR(45) |
| fact\_hapiness\_ analysis | hapiness\_analysis\_id | INT | hapiness\_analysis\_id |
| happiness\_id | INT |
| climat\_id | INT |
| event\_id | INT |
| date\_id | INT |
| country\_id | INT |
| life\_ladder | DECIMAL(10,3) |
| gdp\_per\_capita | DECIMAL(10,3) |
| social\_support | DECIMAL(10,3) |
| life\_expancy | DECIMAL(10,3) |
| freedom\_choice | DECIMAL(10,3) |
| generocity | DECIMAL(10,3) |
| corruption | DECIMAL(10,3) |
| positive\_affect | DECIMAL(10,3) |
| negative affect | DECIMAL(10,3) |

В результаті була спроектована схема сховища даних, яка зображена на рисунку 3.1. Дана модель дозволяє описувати відповідні бізнес-процеси згідно предметній області. Для представлення даних була вибрана багатовимірна модель зі схемою “зірка”.



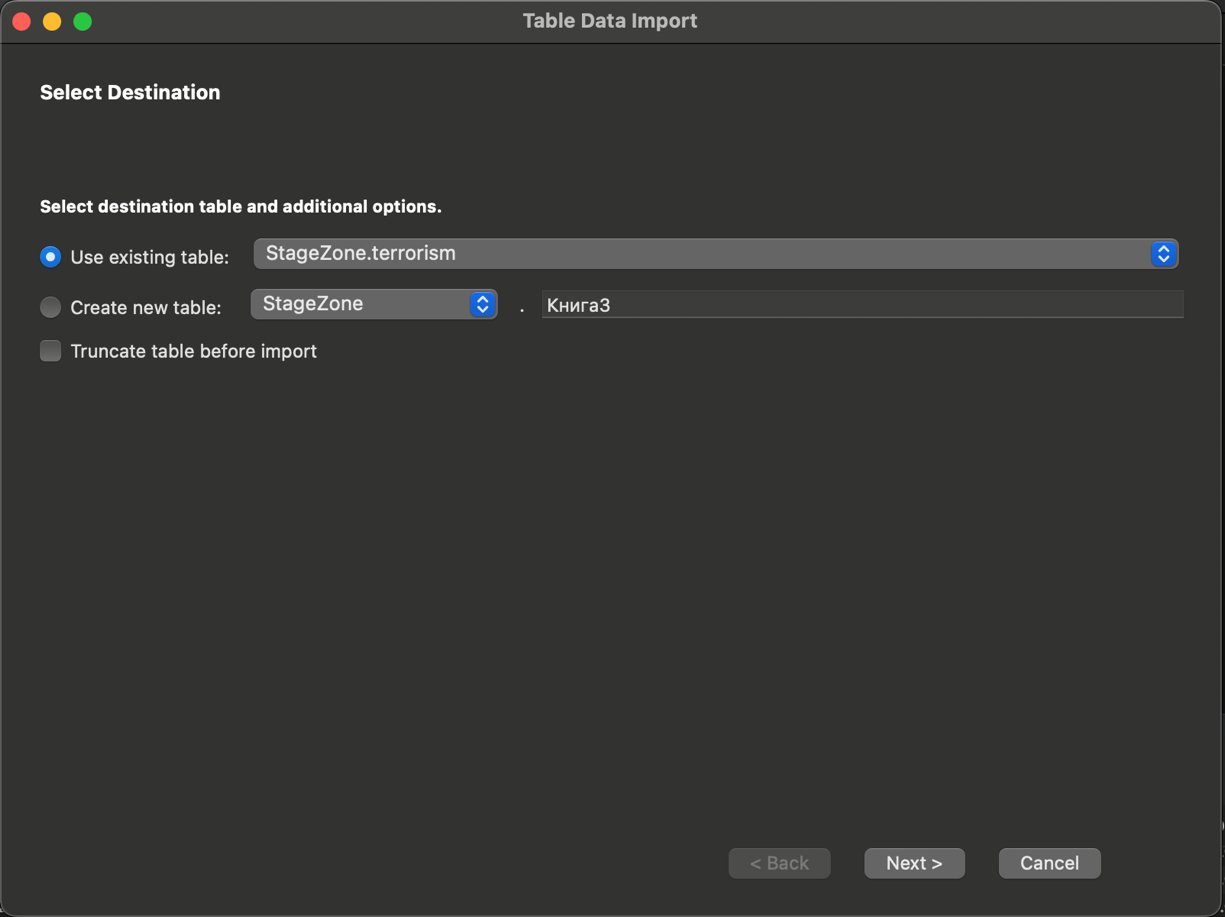
*Рисунок 3.1 – Сховище за типом «зірка»*

Скрипти створення основного сховища за типом «зірка» знаходяться у додатку Б.

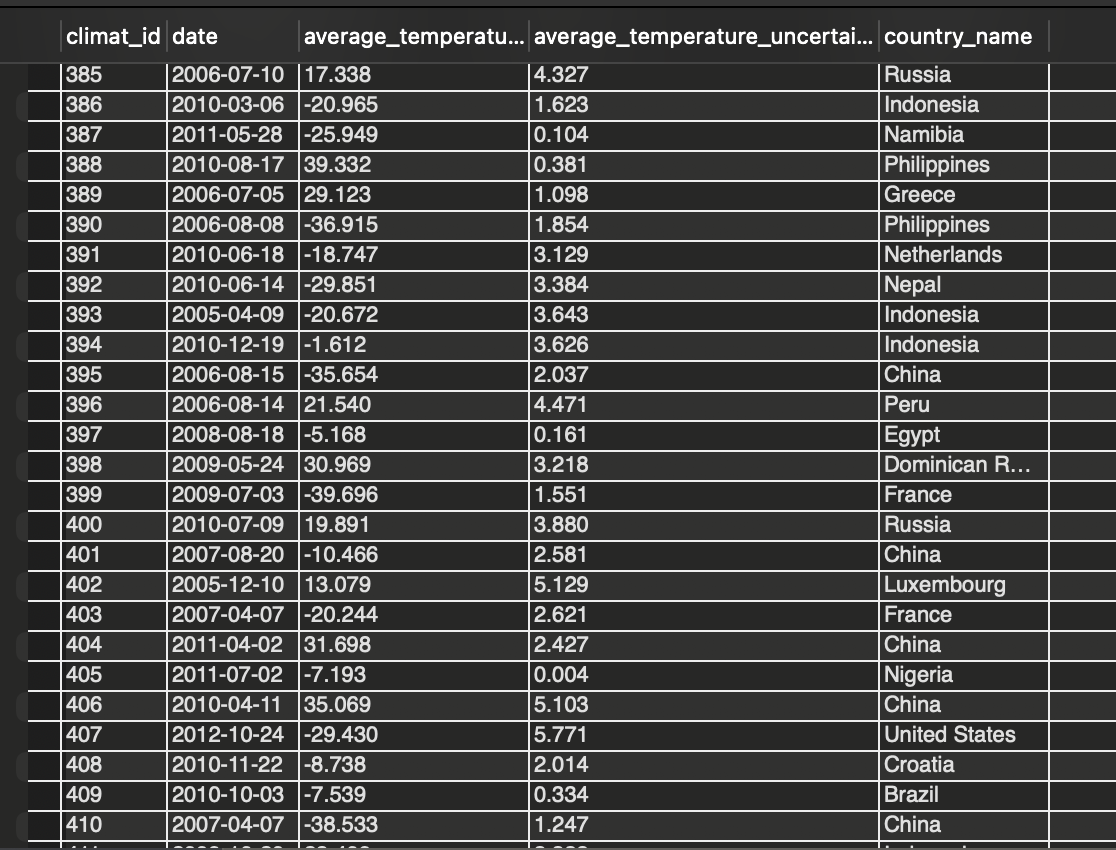
1. ETL засоби

Скрипти ETL-процедур знаходяться у додатку В.

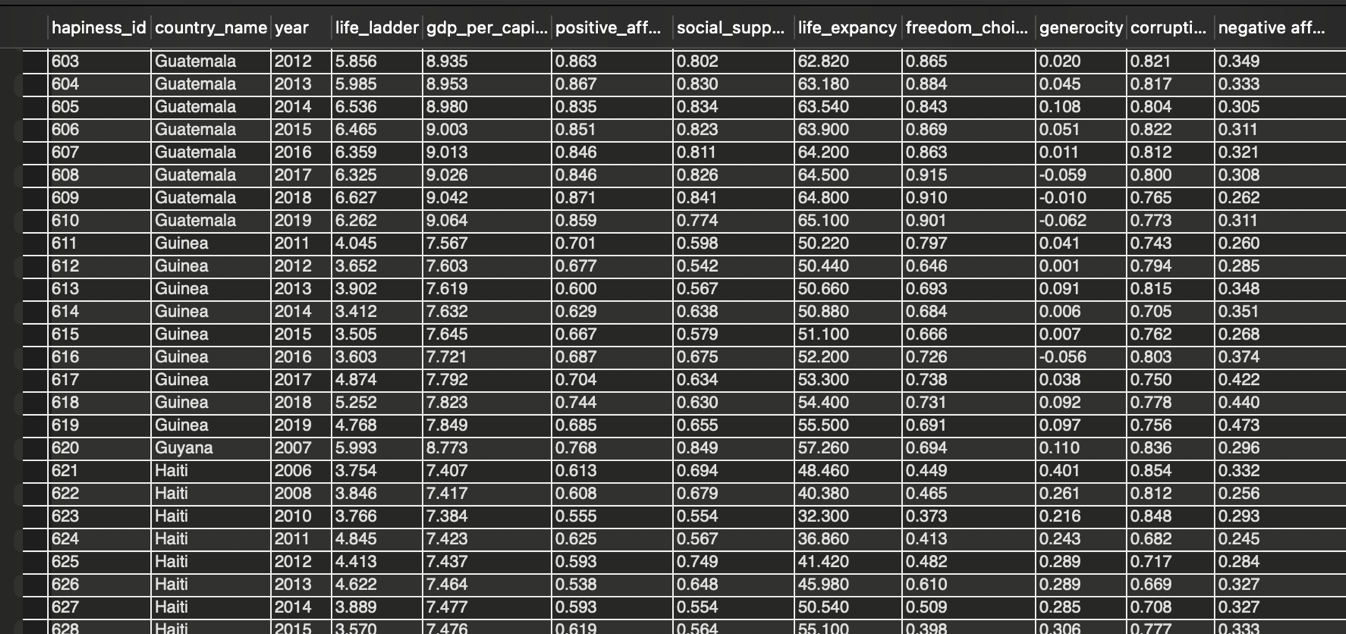
4.1. Завантаження даних до Stage зони



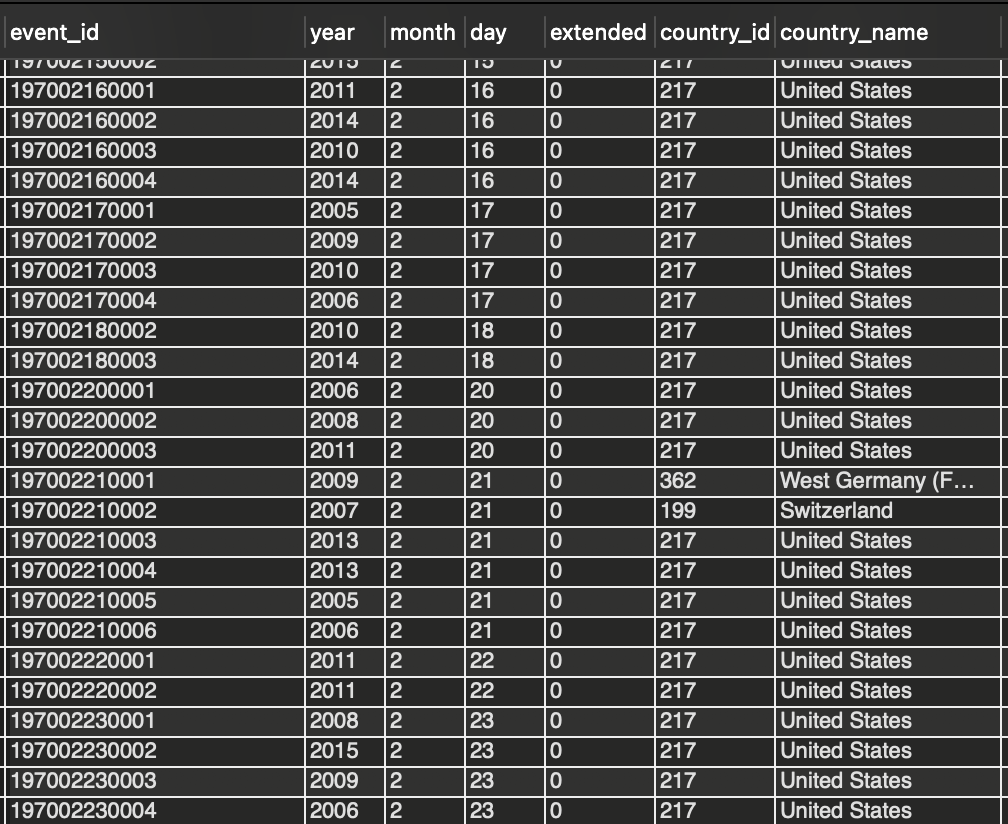
*Рисунок 4.1.1 – Імпорт даних*



*Рисунок 4.1.2 – Таблиця Climat*



*Рисунок 4.1.3 – Таблиця Hapiness*



*Рисунок 4.1.4 – Таблиця Terrorism*

4.2. Cтворення набору процедур/функцій для перетворення та завантаження даних до основного сховища.

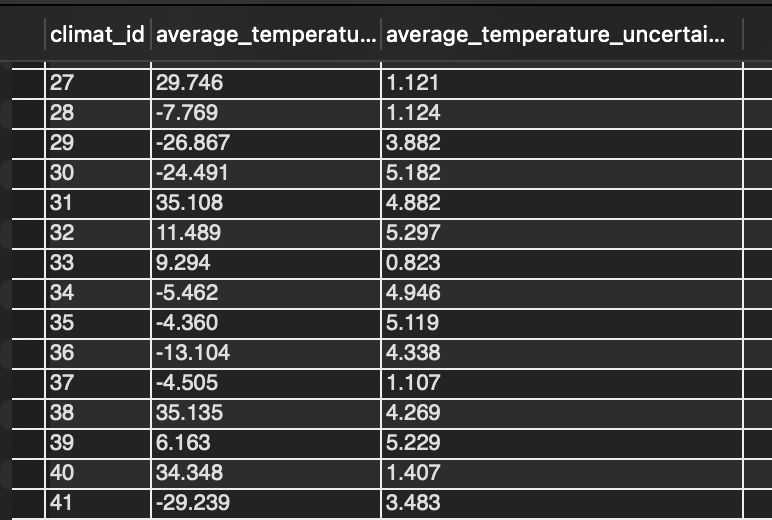
Скрипти ETL-процедур знаходяться у додатку В.

1. Завантаження даних до основного сховища даних.

За допомогую скриптів дані були перенесені у основне сховище для подальшого аналізу даних.



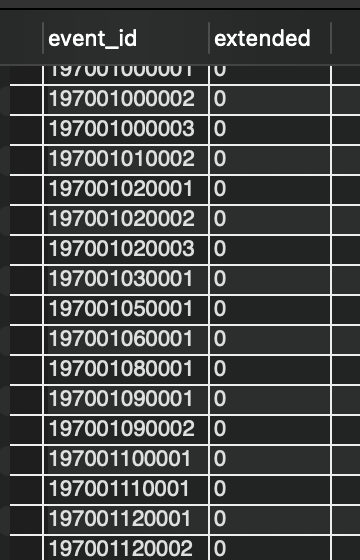
*Рисунок 5.1 – Таблиця dim\_country*



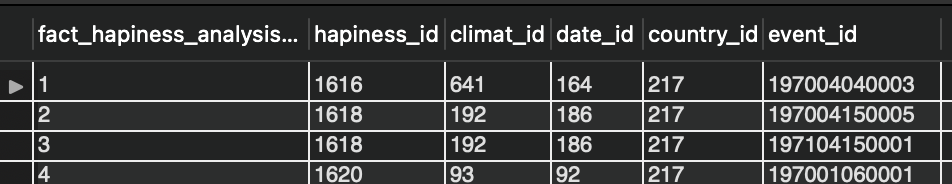
*Рисунок 5.2 – Таблиця dim\_climat*



*Рисунок 5.3 – Таблиця dim\_date*



*Рисунок 5.4 – Таблиця terrorism*



*Рисунок 5.5 – Таблиця fact\_hapiness\_analysis*

Висновок

Ця практична робота дозволила ознайомитися з підходами до створення сховища даних та вивчити процес створення процедур завантаження даних до сховища. Під час виконання було обрано три джерела відкритих даних та створено модель основного сховища за типом "зірка". Було спроектовано модель Stage зони для ETL процесів та створено набір процедур/функцій для перетворення та завантаження даних до основного сховища. Усі етапи роботи було виконано успішно, що дозволяє використовувати сховище даних для подальшого аналізу та використання даних з обраних джерел.

**Додаток А**

-- -----------------------------------------------------

-- Table Hapiness

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS Hapiness (

`hapiness\_id` INT NOT NULL AUTO\_INCREMENT,

`country\_name` VARCHAR(45) NULL,

`year` INT NULL,

`life\_ladder` DECIMAL(10,3) NULL,

`gdp\_per\_capita` DECIMAL(10,3) NULL,

`positive\_affect` DECIMAL(10,3) NULL,

`social\_support` DECIMAL(10,3) NULL,

`life\_expancy` DECIMAL(10,3) NULL,

`freedom\_choice` DECIMAL(10,3) NULL,

`generocity` DECIMAL(10,3) NULL,

`corruption` DECIMAL(10,3) NULL,

`negative affect` DECIMAL(10,3) NULL,

PRIMARY KEY (`hapiness\_id`))

ENGINE = InnoDB;

-- -----------------------------------------------------

-- Table Climat

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS Climat (

`climat\_id` INT NOT NULL AUTO\_INCREMENT,

`date` DATE NULL,

`average\_temperature` DECIMAL(10,3) NULL,

`average\_temperature\_uncertainty` DECIMAL(10,3) NULL,

`country\_name` VARCHAR(45) NULL,

PRIMARY KEY (`climat\_id`))

ENGINE = InnoDB;

-- -----------------------------------------------------

-- Table Terrorism

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS Terrorism (

`event\_id` BIGINT NOT NULL,

`year` INT NULL,

`month` INT NULL,

`day` INT NULL,

`extended` TINYINT(2) NULL,

`country\_id` INT NULL,

`country\_name` VARCHAR(45) NULL,

PRIMARY KEY (`event\_id`))

ENGINE = InnoDB;

**Додаток Б**

-- -----------------------------------------------------

-- Table dim\_climat

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS dim\_climat (

`climat\_id` INT NOT NULL AUTO\_INCREMENT,

`average\_temperature` DECIMAL(10,3) NULL,

`average\_temperature\_uncertainty` DECIMAL(10,3) NULL,

PRIMARY KEY (`climat\_id`))

ENGINE = InnoDB;

-- -----------------------------------------------------

-- Table dim\_terrorism

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS dim\_terrorism (

`event\_id` INT NOT NULL AUTO\_INCREMENT,

`event\_name` VARCHAR(45) NULL,

`extended` TINYINT(2) NULL,

PRIMARY KEY (`event\_id`))

ENGINE = InnoDB;

-- -----------------------------------------------------

-- Table dim\_date

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS dim\_date (

`date\_id` INT NOT NULL AUTO\_INCREMENT,

`year` INT NULL,

`month` INT NULL,

`day` INT NULL,

PRIMARY KEY (`date\_id`))

ENGINE = InnoDB;

-- -----------------------------------------------------

-- Table dim\_country

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS dim\_country (

`country\_id` INT NOT NULL AUTO\_INCREMENT,

`country\_code` INT NULL,

`country\_name` VARCHAR(45) NULL,

PRIMARY KEY (`country\_id`))

ENGINE = InnoDB;

-- -----------------------------------------------------

-- Table fact\_hapiness\_analysis

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS fact\_hapiness\_analysis (

`fact\_hapiness\_analysis\_id` INT NOT NULL AUTO\_INCREMENT,

`climat\_id` INT NULL,

`date\_id` INT NOT NULL,

`country\_id` INT NOT NULL,

`event\_id` INT NULL,

`life\_ladder` DECIMAL(10,3) NULL,

`gdp\_per\_capita` DECIMAL(10,3) NULL,

`positive\_affect` DECIMAL(10,3) NULL,

`social\_support` DECIMAL(10,3) NULL,

`life\_expancy` DECIMAL(10,3) NULL,

`freedom\_choice` DECIMAL(10,3) NULL,

`generocity` DECIMAL(10,3) NULL,

`corruption` DECIMAL(10,3) NULL,

`negative\_affect` DECIMAL(10,3) NULL,

PRIMARY KEY (`fact\_hapiness\_analysis\_id`),

INDEX `fk1\_idx` (`climat\_id` ASC) VISIBLE,

INDEX `fk2\_idx` (`date\_id` ASC) VISIBLE,

INDEX `fk3\_idx` (`country\_id` ASC) VISIBLE,

INDEX `fk4\_idx` (`event\_id` ASC) VISIBLE,

CONSTRAINT `fk1`

FOREIGN KEY (`climat\_id`)

REFERENCES dim\_climat (`climat\_id`)

ON DELETE NO ACTION

ON UPDATE NO ACTION,

CONSTRAINT `fk2`

FOREIGN KEY (`date\_id`)

REFERENCES dim\_date (`date\_id`)

ON DELETE NO ACTION

ON UPDATE NO ACTION,

CONSTRAINT `fk3`

FOREIGN KEY (`country\_id`)

REFERENCES dim\_country (`country\_id`)

ON DELETE NO ACTION

ON UPDATE NO ACTION,

CONSTRAINT `fk4`

FOREIGN KEY (`event\_id`)

REFERENCES dim\_terrorism (`event\_id`)

ON DELETE NO ACTION

ON UPDATE NO ACTION)

ENGINE = InnoDB;

**Додаток В**

-- ----------------------------------------------------

-- Table dim\_climat

-- -----------------------------------------------------

INSERT INTO DataWarehouse.dim\_climat (average\_temperature, average\_temperature\_uncertainty)

SELECT

ROUND(average\_temperature, 3) as average\_temperature,

ROUND(average\_temperature\_uncertainty, 3) as average\_temperature\_uncertainty

FROM StageZone.Climat;

-- ----------------------------------------------------

-- Table dim\_terrorism

-- -----------------------------------------------------

INSERT INTO DataWarehouse.dim\_terrorism (event\_name, extended)

SELECT event\_id, extended

FROM StageZone.Terrorism;

-- ----------------------------------------------------

-- Table dim\_country

-- -----------------------------------------------------

INSERT INTO DataWarehouse.dim\_country (country\_code, country\_name)

SELECT country\_id, country\_name FROM StageZone.Terrorism

UNION

SELECT NULL, country\_name FROM StageZone.Hapiness

UNION

SELECT NULL, country\_name FROM StageZone.Climat

WHERE NOT EXISTS (

SELECT \* FROM DataWarehouse.dim\_country

WHERE DataWarehouse.dim\_country.country\_name = country\_name

);

-- ----------------------------------------------------

-- Table dim\_date

-- -----------------------------------------------------

INSERT IGNORE INTO DataWarehouse.dim\_date (year, month, day)

SELECT DISTINCT YEAR(date) AS year, MONTH(date) AS month, DAY(date) AS day

FROM StageZone.Climat

UNION

SELECT DISTINCT year, NULL, NULL FROM StageZone.Hapiness

UNION

SELECT DISTINCT year, month, day FROM StageZone.Terrorism

WHERE NOT EXISTS (

SELECT \* FROM DataWarehouse.dim\_date

WHERE (DataWarehouse.dim\_date.year = year AND DataWarehouse.dim\_date.month = month AND DataWarehouse.dim\_date.day = day)

);

-- ----------------------------------------------------

-- Table fact\_hapiness\_analysis

-- -----------------------------------------------------

INSERT INTO fact\_hapiness\_analysis

(climat\_id, date\_id, country\_id, event\_id, life\_ladder, gdp\_per\_capita, positive\_affect, social\_support, life\_expancy, freedom\_choice, generocity, corruption, negative\_affect)

SELECT

CL.climat\_id,

D.date\_id,

C.country\_id,

T.event\_id,

SH.life\_ladder,

SH.gdp\_per\_capita,

SH.positive\_affect,

SH.social\_support,

SH.life\_expancy,

SH.freedom\_choice,

SH.generocity,

SH.corruption,

SH.negative\_affect

FROM StageZone.Hapiness SH

JOIN StageZone.Terrorism ST ON ST.country\_name = SH.country\_name

JOIN StageZone.Climat SCL ON SCL.country\_name = ST.country\_name

JOIN dim\_country C ON C.country\_name = SH.country\_name

JOIN dim\_date D ON ST.year = SH.year

AND ST.year = YEAR(SCL.date)

AND ST.month = MONTH(SCL.date)

AND ST.day = DAY(SCL.date)

JOIN dim\_terrorism T ON T.extended = ST.extended

AND T.event\_name = ST.event\_id

JOIN dim\_climat CL ON CL.average\_temperature = SCL.average\_temperature

AND CL.average\_temperature\_uncertainty = SCL.average\_temperature\_uncertainty

WHERE CONCAT(SH.hapiness\_id, CL.climat\_id, D.date\_id, C.country\_id, T.event\_id)

NOT IN (SELECT CONCAT(hapiness\_id, climat\_id, date\_id, country\_id, event\_id)

FROM fact\_hapiness\_analysis);