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

2. 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.

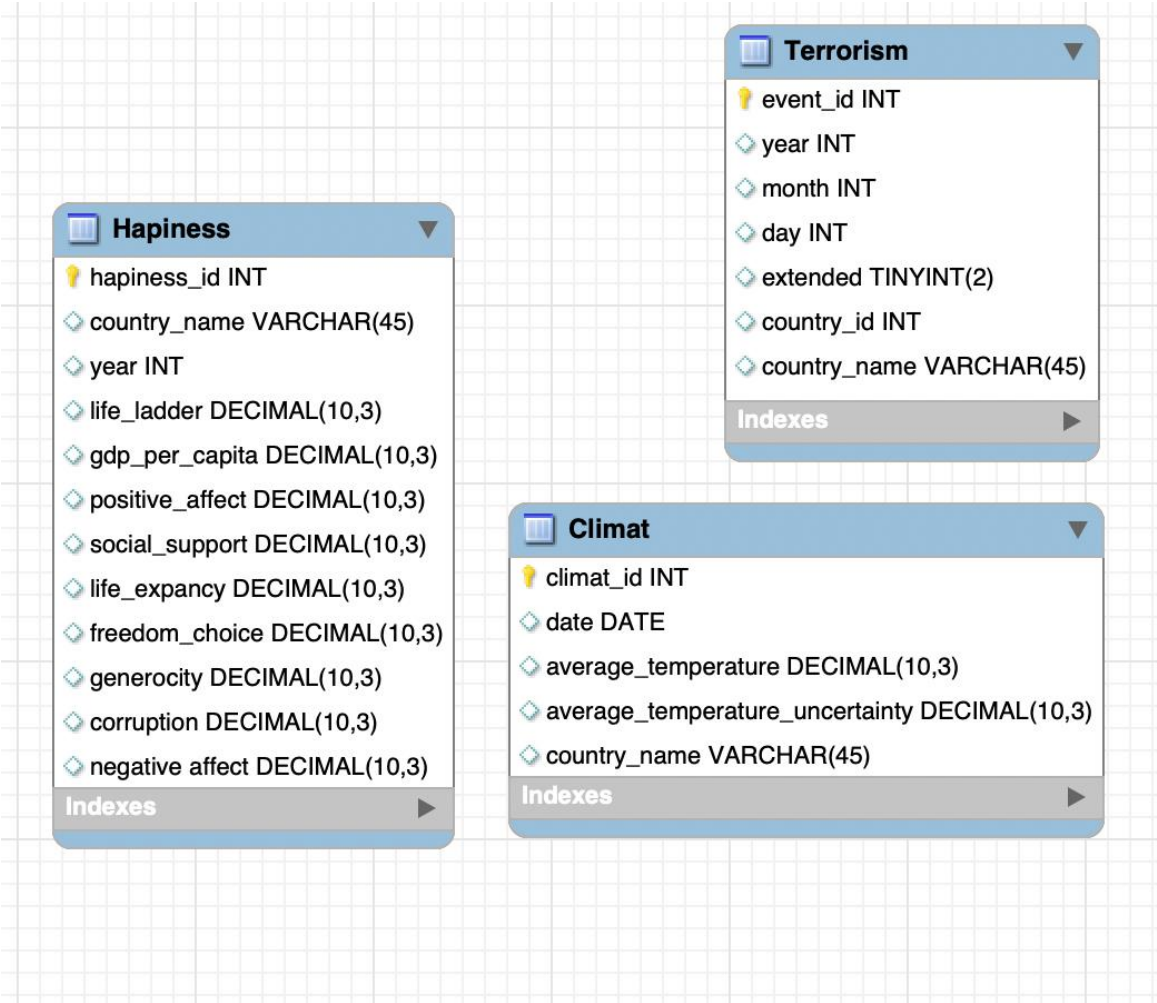


Рисунок 2.1 – Stage зона для ETL процесів

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.

3. 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”.

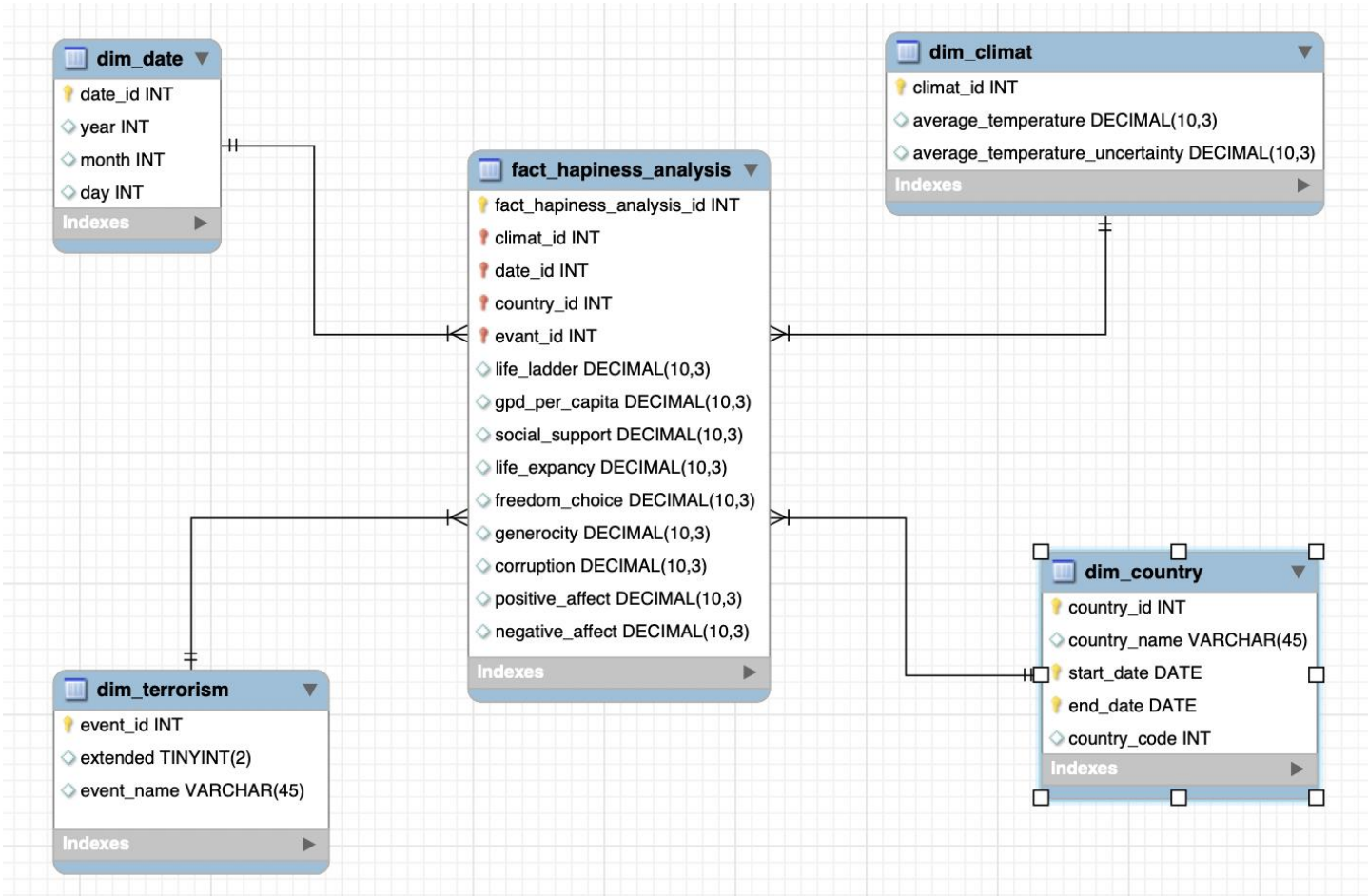


Рисунок 3.1 – Сховище за типом «зірка»

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

4. ETL засоби

Скрипти ETL-процедур знаходяться у додатку В.

4.1. Завантаження даних до Stage зони

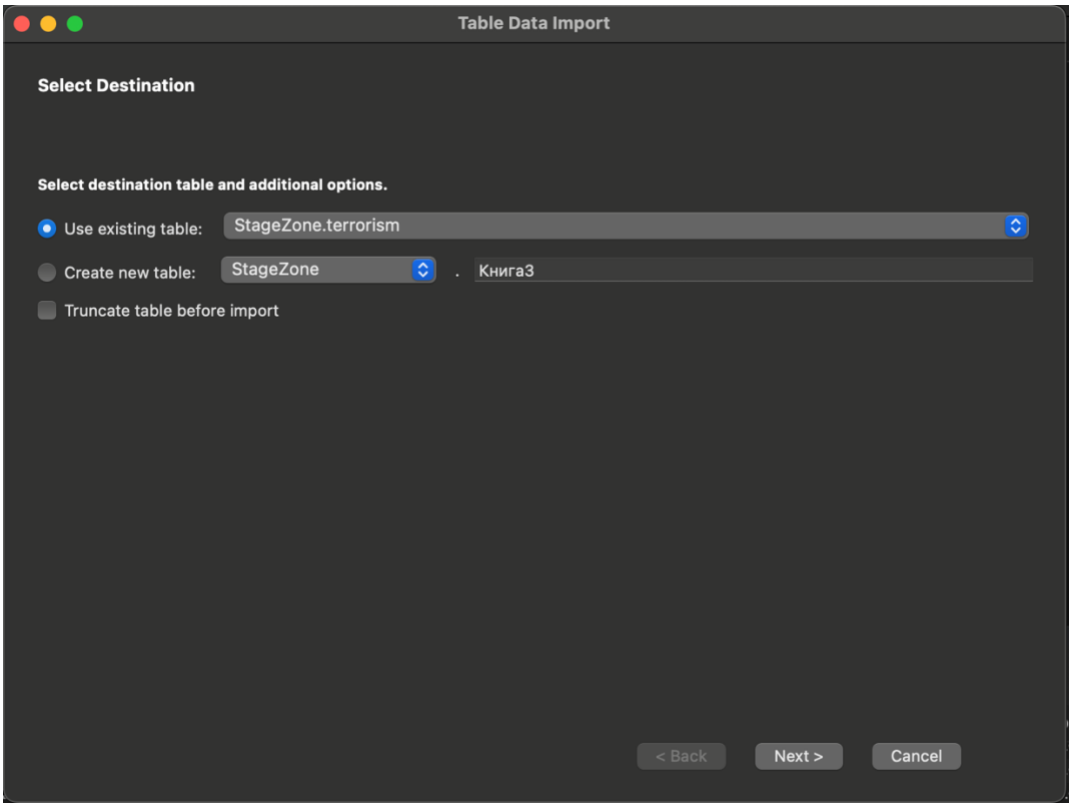


Рисунок 4.1.1 – Імпорт даних

climat_id	date	average_temperatu...	average_temperature_uncertai...	country_name
385	2006-07-10	17.338	4.327	Russia
386	2010-03-06	-20.965	1.623	Indonesia
387	2011-05-28	-25.949	0.104	Namibia
388	2010-08-17	39.332	0.381	Philippines
389	2006-07-05	29.123	1.098	Greece
390	2006-08-08	-36.915	1.854	Philippines
391	2010-06-18	-18.747	3.129	Netherlands
392	2010-06-14	-29.851	3.384	Nepal
393	2005-04-09	-20.672	3.643	Indonesia
394	2010-12-19	-1.612	3.626	Indonesia
395	2006-08-15	-35.654	2.037	China
396	2006-08-14	21.540	4.471	Peru
397	2008-08-18	-5.168	0.161	Egypt
398	2009-05-24	30.969	3.218	Dominican R...
399	2009-07-03	-39.696	1.551	France
400	2010-07-09	19.891	3.880	Russia
401	2007-08-20	-10.466	2.581	China
402	2005-12-10	13.079	5.129	Luxembourg
403	2007-04-07	-20.244	2.621	France
404	2011-04-02	31.698	2.427	China
405	2011-07-02	-7.193	0.004	Nigeria
406	2010-04-11	35.069	5.103	China
407	2012-10-24	-29.430	5.771	United States
408	2010-11-22	-8.738	2.014	Croatia
409	2010-10-03	-7.539	0.334	Brazil
410	2007-04-07	-38.533	1.247	China

Рисунок 4.1.2 – Таблиця Climat

hapiness_id	country_name	year	life_ladder	gdp_per_capi...	positive_aff...	social_supp...	life_expancy	freedom_choi...	generocity	corrupti...	negative aff...
603	Guatemala	2012	5.856	8.935	0.863	0.802	62.820	0.865	0.020	0.821	0.349
604	Guatemala	2013	5.985	8.953	0.867	0.830	63.180	0.884	0.045	0.817	0.333
605	Guatemala	2014	6.536	8.980	0.835	0.834	63.540	0.843	0.108	0.804	0.305
606	Guatemala	2015	6.465	9.003	0.851	0.823	63.900	0.869	0.051	0.822	0.311
607	Guatemala	2016	6.359	9.013	0.846	0.811	64.200	0.863	0.011	0.812	0.321
608	Guatemala	2017	6.325	9.026	0.846	0.826	64.500	0.915	-0.059	0.800	0.308
609	Guatemala	2018	6.627	9.042	0.871	0.841	64.800	0.910	-0.010	0.765	0.262
610	Guatemala	2019	6.262	9.064	0.859	0.774	65.100	0.901	-0.062	0.773	0.311
611	Guinea	2011	4.045	7.567	0.701	0.598	50.220	0.797	0.041	0.743	0.260
612	Guinea	2012	3.652	7.603	0.677	0.542	50.440	0.646	0.001	0.794	0.285
613	Guinea	2013	3.902	7.619	0.600	0.567	50.660	0.693	0.091	0.815	0.348
614	Guinea	2014	3.412	7.632	0.629	0.638	50.880	0.684	0.006	0.705	0.351
615	Guinea	2015	3.505	7.645	0.667	0.579	51.100	0.666	0.007	0.762	0.268
616	Guinea	2016	3.603	7.721	0.687	0.675	52.200	0.726	-0.056	0.803	0.374
617	Guinea	2017	4.874	7.792	0.704	0.634	53.300	0.738	0.038	0.750	0.422
618	Guinea	2018	5.252	7.823	0.744	0.630	54.400	0.731	0.092	0.778	0.440
619	Guinea	2019	4.768	7.849	0.685	0.655	55.500	0.691	0.097	0.756	0.473
620	Guyana	2007	5.993	8.773	0.768	0.849	57.260	0.694	0.110	0.836	0.296
621	Haiti	2006	3.754	7.407	0.613	0.694	48.460	0.449	0.401	0.854	0.332
622	Haiti	2008	3.846	7.417	0.608	0.679	40.380	0.465	0.261	0.812	0.256
623	Haiti	2010	3.766	7.384	0.555	0.554	32.300	0.373	0.216	0.848	0.293
624	Haiti	2011	4.845	7.423	0.625	0.567	36.860	0.413	0.243	0.682	0.245
625	Haiti	2012	4.413	7.437	0.593	0.749	41.420	0.482	0.289	0.717	0.284
626	Haiti	2013	4.622	7.464	0.538	0.648	45.980	0.610	0.289	0.669	0.327
627	Haiti	2014	3.889	7.477	0.593	0.554	50.540	0.509	0.285	0.708	0.327
628	Haiti	2015	3.570	7.476	0.619	0.564	55.100	0.398	0.306	0.777	0.333

Рисунок 4.1.3 – Таблиця Hapiness

event_id	year	month	day	extended	country_id	country_name
197002160001	2011	2	16	0	217	United States
197002160002	2014	2	16	0	217	United States
197002160003	2010	2	16	0	217	United States
197002160004	2014	2	16	0	217	United States
197002170001	2005	2	17	0	217	United States
197002170002	2009	2	17	0	217	United States
197002170003	2010	2	17	0	217	United States
197002170004	2006	2	17	0	217	United States
197002180002	2010	2	18	0	217	United States
197002180003	2014	2	18	0	217	United States
197002200001	2006	2	20	0	217	United States
197002200002	2008	2	20	0	217	United States
197002200003	2011	2	20	0	217	United States
197002210001	2009	2	21	0	362	West Germany (F…
197002210002	2007	2	21	0	199	Switzerland
197002210003	2013	2	21	0	217	United States
197002210004	2013	2	21	0	217	United States
197002210005	2005	2	21	0	217	United States
197002210006	2006	2	21	0	217	United States
197002220001	2011	2	22	0	217	United States
197002220002	2011	2	22	0	217	United States
197002230001	2008	2	23	0	217	United States
197002230002	2015	2	23	0	217	United States
197002230003	2009	2	23	0	217	United States
197002230004	2006	2	23	0	217	United States

Рисунок 4.1.4 – Таблиця Terrorism

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

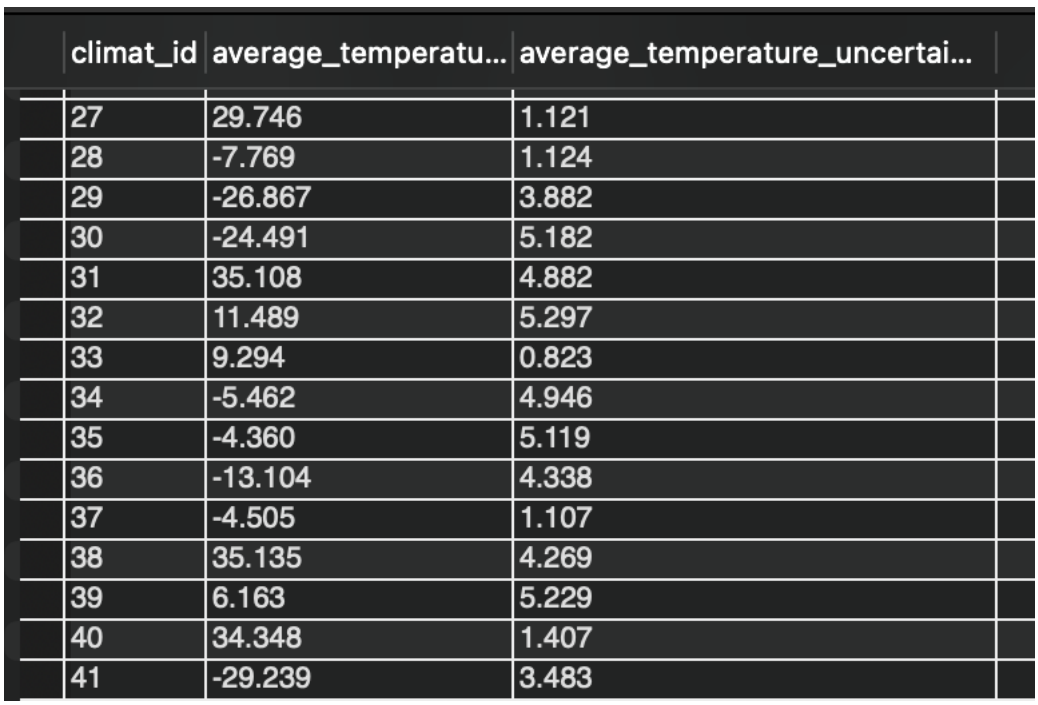
Скрипти ETL-процедур знаходяться у додатку В.

5. Завантаження даних до основного сховища даних.

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

country_id	country_name
11	Argentina
14	Australia
21	Belgium
26	Bolivia
30	Brazil
36	Cambodia
38	Canada
45	Colombia
49	Costa Rica
58	Dominican Republic
60	Egypt
65	Ethiopia
78	Greece

Рисунок 5.1 – Таблиця dim_country



climat_id	average_temperatu...	average_temperature_uncertai...
27	29.746	1.121
28	-7.769	1.124
29	-26.867	3.882
30	-24.491	5.182
31	35.108	4.882
32	11.489	5.297
33	9.294	0.823
34	-5.462	4.946
35	-4.360	5.119
36	-13.104	4.338
37	-4.505	1.107
38	35.135	4.269
39	6.163	5.229
40	34.348	1.407
41	-29.239	3.483

Рисунок 5.2 – Таблиця dim_climat

date_id	year	month	day
3	2011	4	5
4	2011	3	19
5	2008	9	22
6	2012	12	7
7	2008	2	7
8	2005	11	26
9	2009	1	7
10	2007	7	9
11	2009	8	20
12	2006	3	25
13	2011	9	11
14	2012	3	28
15	2011	11	14

Рисунок 5.3 – Таблиця dim_date

event_id	extended
197001000001	0
197001000002	0
197001000003	0
197001010002	0
197001020001	0
197001020002	0
197001020003	0
197001030001	0
197001050001	0
197001060001	0
197001080001	0
197001090001	0
197001090002	0
197001100001	0
197001110001	0
197001120001	0
197001120002	0

Рисунок 5.4 – Таблиця terrorism

fact_hapiness_analysis...	hapiness_id	climat_id	date_id	country_id	event_id
1	1616	641	164	217	197004040003
2	1618	192	186	217	197004150005
3	1618	192	186	217	197104150001
4	1620	93	92	217	197001060001

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);
```