МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО”

Факультет прикладної математики

Кафедра програмного забезпечення комп’ютерних систем

КУРСОВИЙ ПРОЕКТ

з дисципліни “Бази даних”

спеціальність 121 – Програмна інженерія

на тему: Система аналізу поширення захворюванності на “Covid-19”

Студент
групи КП-83 Ландо Максим Юрійович _____
(підпис)

Викладач
к.т.н, доцент кафедри СПіСКС **Петрашенко А.В.** _____
(підпис)

Захищено з оцінкою _____

Київ – 2021

Анотація

Метою розробки даного курсового проекту є набуття практичних навичок розробки сучасного програмного забезпечення, що взаємодіє з пост реляційними базами даних, а також здобуття навичок оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. У результаті виконання курсового проекту було опановано навик розробляти програмне забезпечення для пост реляційних баз даних, володіння основами використання СКБД, а також інструментальними засобами аналізу великих обсягів даних.

Темою даного курсового проекту є створення системи аналізу швидкості поширення та тенденцій розповсюдження коронавірусного захворювання “Covid-19”. У документі викладена актуальність та проблематика аналізу великого обсягу даних, аналіз використаного інструментарію (опис мови програмування, використаних бібліотек та СКБД), описана структура бази даних, опис розробленого програмного забезпечення (загальний, опис модулів та основних алгоритмів роботи).

Результатами даного проекту стали діаграми та графіки, що зображають результати дослідження даних щодо поширення “Covid-19”.

ЗМІСТ

Вступ

Під час виконання проекту було створено систему аналізу поширення коронавірусного захворювання “Covid-19”. Всі дані було отримано з датасету European Centre for Disease Prevention and Control

Актуальність

На сьогоднішній день кожен найбільш обговорюваною хворобою є “Covid-19”, в новинах та газетах щодня публікується безліч інформації про дану хворобу. Даною хворобою вже перехворіло значна частина насеення Землі.

Програма працює з даними щодо “Covid-19” за 2020 рік. Програма рахує медіану, середнє значення, тенденції поширення захворювання в різних країнах за різні часові проміжки.

Мета розробки

Створення програмного забезпечення, що забезпечує роботу наведених далі пунктів:

1. Попередня обробка даних: Засоби фільтрації і валідації даних. Розроблено ПЗ, що фільтрує дані, отримані з dataset-у та залишає лише потрібні, виправляє неправильні формати та значення.
2. Основний модуль процесу. Аналіз даних отриманих з бази даних
3. Додатковий модуль. Робота з базою даних, імпорт та отримання даних

Дані для аналізу були взяті із відкритого dataset-у за посиланням https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Аналіз інструментарію для виконання курсового проект

Аналіз СКБД

В процесі виконання цього курсового проекту перед нами встала потреба кешувати та зберігати дані про “Covid-19” між запусками аналізатора. Кожного разу читати дані із CSV - файлів є дуже дорогою операцією, тож було прийняте рішення використати СКБД. В якості СКБД були розглянуті варіанти: PostgreSQL, MongoDB, CassandraDB. З порівняльною характеристикою цих СКБД можна ознайомитися в таблиці 1.

таблиця 1. Порівняльна характеристика СКБД

Критерій порівняння	Назва СКБД		
	MongoDB	PostgreSQL	CassandraDB
Реляційні дані	ні	так	так
Схема даних	динамічна	статична і динамічна	статична і динамічна
Підтримка ієрархічних даних	так	так	ні
Має відкритий вихідний код	так	так	так
Транзакції	ні	так	так
Атомарність операцій	всередині документа	по всій БД	всередині партиції

Мова запитів	JSON/JavaScript	SQL	CQL
Найлегший спосіб масштабування	горизонтальний	вертикальний	горизонтальний
Підтримка шардингів	так	так (важка конфігурація)	так (може зберігати партиції на різних машинах)
Приклад використання	Великі дані (мільярди записів) з великою кількістю паралельних оновлень, де цілісність і узгодженість даних не потрібно.	Транзакційні і операційні програми, вигода яких в нормалізованому формі, об'єднаннях, обмеження даних і підтримки транзакцій.	Багато запитів на запис/читання у одиницю часу, до даних можна задіяти партиціювання за ключем, дані мають лише первинні індекси
Наявність бібліотек для мови програмування Node 12	так	так	так
Підтримка реплікації	так, автоматичне переобрання головного процесу	За принципом master-slave	так, через партиціювання
Засіб збереження та відновлення даних	mongodump	pg_dump	не має окремого доданка, виконується засобами CQL

Форма збереження даних	документы JSON	таблиця	таблиця

За результатами порівнянні цих СКБД було прийнято рішення зупинитися на NoSQL рішеннях. Оскільки вона чудово поєднує в собі переваги неструктурованих баз даних. Крім цього класичним прикладом використання NoSQL СКБД є системи збору та аналізу даних, до яких можна застосувати індексування за первинними та вторинними ключами.

Ця база даних є об’єктно орієнтованою та дозволяє зберігати великі масиви неструктурованих даних. На відміну від SQL баз даних ми можемо зберігати дані у “сирому” об’єктному вигляді, який використовується програмою та є більш близьким за структурою до моделі даних, яку буде використовувати ПЗ написане з використанням мови програмування Python. Це пришвидшить збір, збереження та отримання даних програмним забезпеченням. Оскільки MongoDB є представником NoSQL баз даних, вона не потребує жорсткої схеми даних, що дозволяє пришвидшити процес розробки та зробити його більш гнучким. Окрім цього дана СКБД підтримує горизонтальне масштабування за допомогою шардингу з метою зменшення навантаження на кожен окремий вузол шляхом розподілення навантаження між ними всіма.

Нижче наведено перелік основних переваг:

- Динамічна схема
- Швидкість запису у колекцію
- Швидкість читання із колекції

Обґрунтування вибору мови програмування

Мовою програмування для ПЗ було обрано Python. Оскільки це дуже популярна мова з безліччю бібліотек як для роботи з базами даних, так і з для обробки та аналізу інформації. Вона дозволяє створити гнучку систему, що спрощує розробку та подальше розширення програми

Обґрунтування вибору бібліотек і фреймворків

Використані бібліотеки:

- pymongo – це бібліотека Python, яка дозволяє взаємодіяти з базою даних MongoDB, створювати зв'язок з нею, отримувати та передавати дані.
- pandas – це бібліотекаPython, яка надаєінструменти роботи з датафреймом, різноманітного аналізу та перетворень даних.
- matplotlib – це бібліотека Python, яка призначена для створення різноманітної візуалізації даних (графіки, діаграми тощо).

Структура бази даних

База даних складається з однієї колекції, в якій зберігаються відомості щодо країни, кількості нових випадків та смертей за певний день. Загальна структура документа в базі даних приведена у таблиці 2.

таблиця 2. Опис властивостей документа у базі даних

Назва властивості	Тип	Опис
_id	ObjectId	Ідентифікатор запису
dateRep	String	Повна дата
day	Integer	Номер дня в місяці
month	Integer	Номер місяця в році
year	Integer	Номер року
cases	Integer	Кількість нових випадків
deaths	Integer	Кількість смертей
countriesAndTerritories	String	Повна назва країни
geoId	String	Унікальна назва країни з 2 букв
countryterritoryCode	String	Скорочена назва країни з 3 букв
popData2019	Integer	Кількість населення за 2019 рік
continentExp	String	Назва континенту
Cumulative_number_for_14_days	Float	Середній ріст захворюваності за останні 14 днів

Опис результатів предметної галузі

В результаті виконання курсового проекту було проаналізовано поширення коронавірусного захворювання “Covid-19” за 2020 рік. Було виконано наступне:

- Проаналізована кількість випадків захворювання та смертей за певний період;
- Утворено графіки середніх та медіанних значень, а також тенденції в кожній країні;
- Проаналізовано країни з найбільшими та найменшими кількостями захворювань та смертей за певний період;
- Розроблений консольний інтерфейс для того, щоб користувач міг отримати необхідні графіки, дані, статистику;
- Додана функція резервної копії бази даних.

Висновки

У ході виконання курсового проекту роботи було отримано практичні навички обробки даних та роботи з даними за допомогою мови програмування Python.

Також було отримано нові навички роботи з певними бібліотеками, які спрощують та покращують роботу з Python. Було складено таблиці порівняння баз даних. На основі цих даних було обрано в якості СКБД NoSQL рішення MongoDB.

На основі зібраних даних було побудовано таблиці та графіки. Дані аналізу приведені у Додатку А.

В ході виконання даного курсового проекту було набуто практичних навичок розробки програмного забезпечення, що взаємодіє з NoSQL базою даних, а також були покращені навички оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. Також було набуто навиків в роботі з бібліотеками для побудови графічної частини роботи.

Література

1. PostgreSQL vs MongoDB https://habr.com/post/272735/;
2. Mongodb https://docs.mongodb.com/
3. Matplotlib https://matplotlib.org/
4. Pandas https://pandas.pydata.org/docs/

Додакти

Додаток А

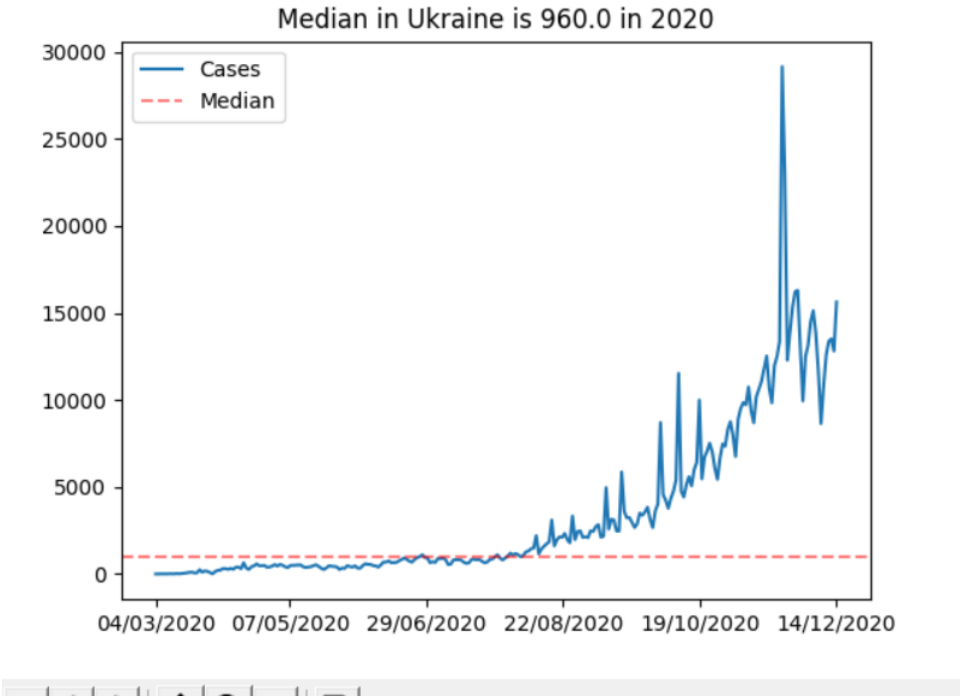

Рис. 1 Медіана захворюваності по країні

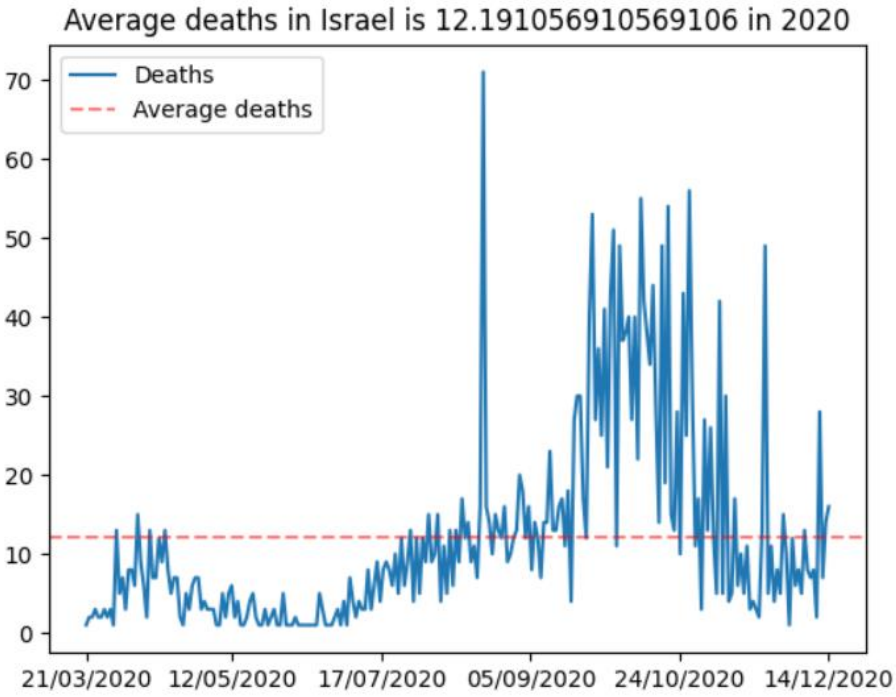

Рис. 2 Середні значення смертності по країні

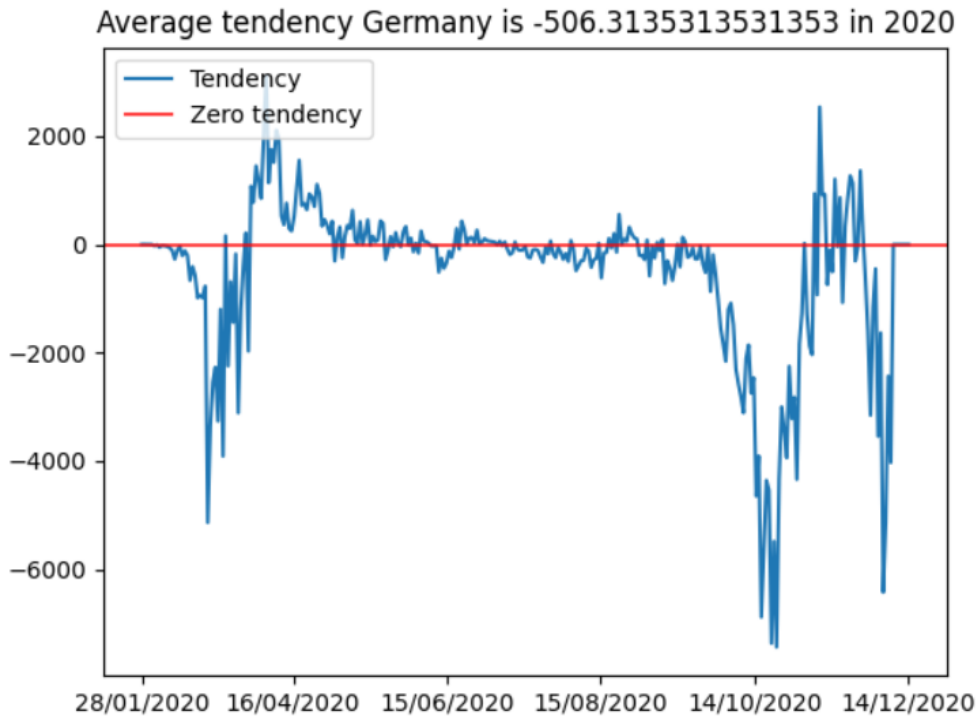

Рис. 3 Тенденції захворюваності по країні

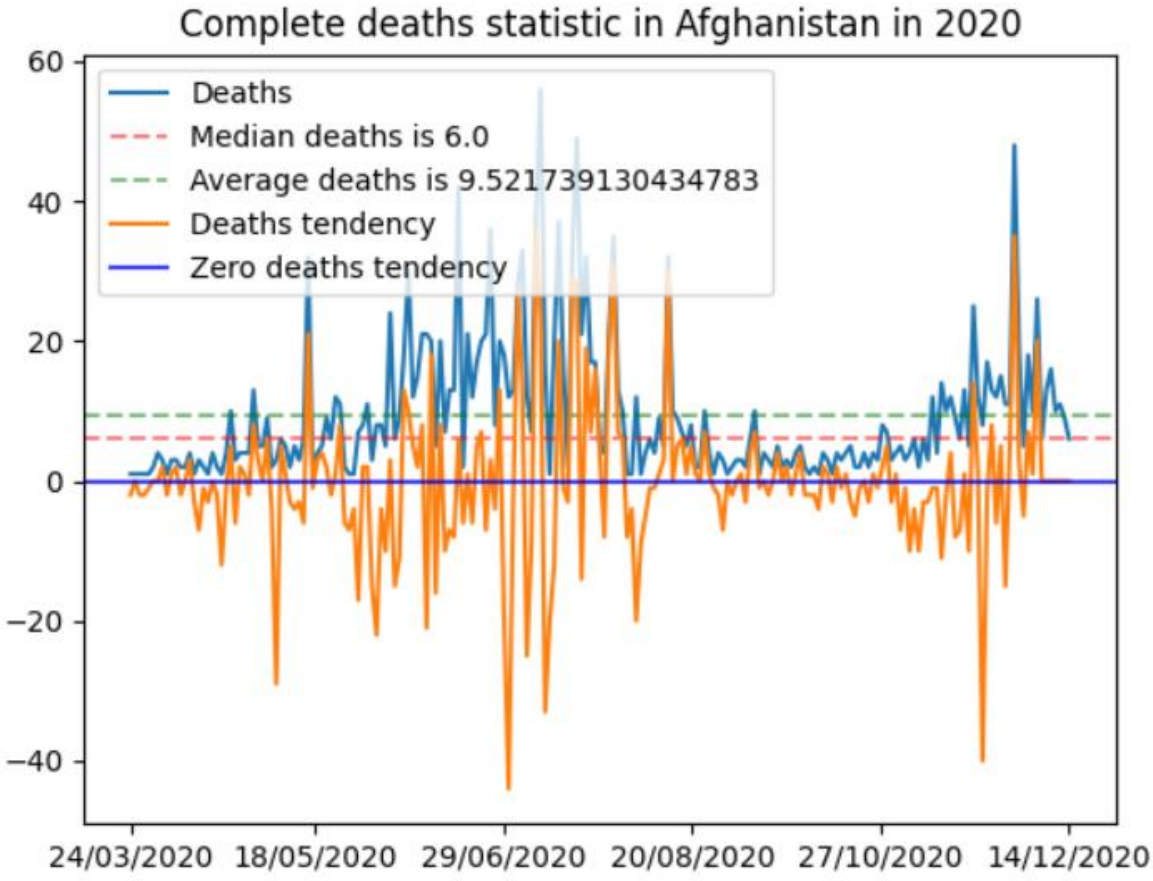

Рис. 4 Загальна статистика смертності по країні

```
Enter start of interval:
10-05-2020
Enter end of interval:
10-08-2020


Best countries by tendency of new cases in period:
                                  cases
countriesAndTerritories
Laos                                  0
Brunei_Darussalam                     0
Cameroon                              0
Somalia                               0
Cayman_Islands                        0
Dominica                              0
Saint_Vincent_and_the_Grenadines      0
Saint_Kitts_and_Nevis                 0
Eritrea                               0
Falkland_Islands_(Malvinas)           0
```

```
Worst countries by tendency of new cases in period:
                           cases
countriesAndTerritories
India                     286282
United_States_of_America  166850
Brazil                     93861
United_Kingdom             64258
Spain                      58392
Argentina                  50097
France                     46884
Russia                     44117
Colombia                   29536
Belgium                    18774
```

Рис. 4-5 Країни з найменшими та найбільшими показниками захворюваності за обраний період

```
6
Enter start of interval:
04-01-2020
Enter end of interval:
11-01-2020


Deaths in period:
              deaths
dateRep
31/10/2020  1157853
31/08/2020  1150533
31/07/2020  1146760
31/05/2020  1140377
30/10/2020  1136367
30/09/2020  1129181
30/08/2020  1122963
30/07/2020  1117265
30/06/2020  1110624
30/05/2020  1106920
```

Рис. 6 Загальна кількість смертей за обраний період