**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**

**ТЕХНІЧНЕ ЗАВДАННЯ**

з дисципліни “Бази даних”

спеціальність 121 – Програмна інженерія

на тему: “Моніторингова система кліматичних показників та якості повітря”

**Студент**

**групи** КП-72 Абу Шамала А.М. \_\_\_\_\_\_\_\_\_\_

(підпис)

**Викладач**

**к.т.н, доцент кафедри**

**СПіСКС Петрашенко А.В.**  \_\_\_\_\_\_\_\_\_\_

(підпис)

Київ – 2020

1. **НАЙМЕНУВАННЯ ТА ГАЛУЗЬ ЗАСТОСУВАННЯ РОЗРОБКИ**

Найменування: моніторингова система кліматичних показників довкілля: збір даних показників температури повітря, вологості повітря, рівня шкідливих речовин у повітрі.

Галузь застосування: збір, аналіз та фільтрація інформації про якість повітря та погодних умов. Передбачення та прогнозування зміни погодних умов та якості повітря.

1. **ДАТА ПОЧАТКУ ТА ЗАКІНЧЕННЯ ПРОЕКТУ**

Дата початку проекту - 26 лютого 2020 року (дата видачі завдання курсового проекту).

Дата закінчення проекту - 23 траваня 2020 року (захист курсового проекту).

1. **МЕТА РОЗРОБКИ**

Метою розробки даного курсового проекту є набуття виконавцем (студентом) практичних навичок розробки сучасного програмного забезпечення, що взаємодіє з постреляційними базами даних, а також здобуття навичок оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. У результаті виконання курсового проекту студент повинен вміти розробляти програмне забезпечення для постреляційних баз даних, володіти основами використання СУБД, а також інструментальними засобами аналізу великих обсягів даних.

1. **ВИМОГИ ДО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ**

* Підсистема попередньої обробки даних містить у собі:

Засоби генерації даних: має бути реалізовано ПЗ для збору інформації про погодні умови та якість повітря, що розміщені на веб-сайтах

<https://www.iqair.com/air-pollution-data-api>

<https://www.saveecobot.com/maps/kyiv> [https://www.wunderground.com/history/daily/ua/kyiv/UKKK/](https://www.wunderground.com/history/daily/ua/kyiv/UKKK/date/2019-5-21)

* Засоби фільтрації та валідації даних:

Доповнення ПЗ з попереднього пункту функціоналом фільтрації та перевірки даних на коректність, відсіювання дублікацій та невідповідних даних.

* База даних:

MongoDB

* Засоби реплікації даних:

оскільки для використання у даній розробці була обрана нереляційна база даних MongoDB, то як засіб реплікації використовуватимемо реплісети (Replica Sets).

* Засоби масштабування:

шардинг (sharding), який використовує MongoDB для обробки великої кількості даних.

* Засоби аналізу даних:

Pandas – бібліотека, яка використовуватиметься як надбудова до NumPy для структуризації роботи із масивами даних.

NumPy – бібліотека для роботи із великими масивами даних.

Matpotlib – бібліотека для візуалізації даних у вигляді 2D і 3D графіків.

* Задачі аналізу даних:

Структурувати усі отримані дані з різних веб-ресурсів про оголошення: час, температура, відносна вологість, атмосферний тиск, забруднення повітря (AQI PM2.5), швидкість вітру

Здійснити валідацію даних та прибрати зайву та не відповідну інформацію із усіх масивів даних.

Об’єднати інформацію та заповнити пусті місця у даних.

Провести кореляцію по всіх даних.

У місцях, де найбільша кореляція, здійснити більш детальний аналіз.

Знайти зв'язок між показником забрудненням, швидкістю вітру та відносною вологістю.

* Засоби резервування та відновлення даних:

Передбачені при використанні Replica Sets у MongoDb.

1. **ОБҐРУНТУВАННЯ ВИБОРУ СУБД**

В даному курсовому проекті використовується MongoDB. 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 СУБД підтримує горизонтальне масштабування за допомогою шардингу з метою зменшення навантаження на кожен окремий вузол шляхом розподілення навантаження між ними всіма.

1. **ВИМОГИ ДО ІНТЕРФЕЙСУ КОРИСТУВАЧА**

Інтерфейс користувача буде консольним. Інтерфейсу користувача буде використовуватися для запуску на виконання ПЗ, його налаштування та передачі параметрів для збору та аналізу даних, генерації звітної інформації (графіків та діаграм) у вигляді збережених на диск файлів - зображень. Звітна інформація стосується візуалізації роботи засобів аналізу даних.

1. **ВИБІР ЗАСОБІВ РОЗРОБКИ**

Мова програмування – Python 3.6.9. Дана мова програмування відносно проста та за думкою фахівців, які займаються розробкою ПЗ у сфері науки про дані, є найкращим рішенням для даного роду задач. Окрім цього для цієї мови існує велика кількість бібліотек для аналізу даних. Зокрема:

* sklearn — бібліотека алгоритмів машинного навчання, використовується для класифікації досліджених даних;
* pandas — бібліотека для обробки та аналізу даних, використовується для первинної обробки даних;
* matplotlib — бібліотека для графічного представлення даних, а саме для побудови графіків;
* numpy — математична бібліотека з підтриманням багатовимірних масивів;

1. **ЕТАПИ РОЗРОБКИ**

|  |  |  |
| --- | --- | --- |
| **№** | **Назва етапів розроблення** | **Термін виконання** |
| 1 | Затвердження теми курсової роботи. Опрацювання відповідної літератури. Розроблення та узгодження технічного завдання. | 12.03.2020 |
| 2 | Аналіз постановки задачі | 20.03.2020 |
| 3 | Розробка засобів генерації даних. | 15.04.2020 |
| 4 | Додавання засобів фільтрації та валідації даних. | 20.04.2020 |
| 5 | Реалізація зберігання, реплікації та масштабування інформації розробленої моніторингової системи. | 25.04.2020 |
| 6 | Додавання засобів аналізу даних (реалізацію алгоритмів буде запозичено у великих бібліотеках аналізу даних). | 30.04.2020 |
| 7 | Додавання засобів резервування та відновлення даних (призначені для оперативного та пакетного збереження фрагментів та усієї бази даних з можливістю її відновлення з урахуванням необхідності підключення додаткового комп’ютера як елемента горизонтального масштабування). | 05.05.2020 |
| 8 | Тестування програми | 10.05.2020 |
| 9 | Аналіз результатів. Підготовка матеріалів курсового проекту та оформлення пояснювальної записки | 15.05.2020 |
| 10 | Захист курсової роботи | 23.05.2020 |