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

Факультет прикладної математики
Кафедра програмного забезпечення комп’ютерних систем

Курсова робота

з дисципліни “Бази даних”

тема “Моніторингова система заробітних плат працівників ІТ індустрії”

Виконав
студент ІІІ курсу
групи КП-83
Мричко Богдан Тарасович
(*прізвище, ім’я, по батькові*)

Приняв
“____” “_____” 2021 р.
викладачем, к.т.н, доцентом кафедри
Радченко Костянтин Олександрович
(*прізвище, ім’я, по батькові*)

Захищено з оцінкою _____

Київ 2021

Анотація

Метою даної курсової роботи є розробка системи опрацювання та аналізу даних, анонімної звітності працівників сфери IT, щодо своїх прибутків, технологій що використовуються у професійній діяльності, можливостей для кар’єрного росту, досвіду робота та інше.

Виконання даної курсової роботи потребує знань та навичок при роботі зі спеціалізованими програмними бібліотеками, постреляційною базою даних та математичного підґрунтя для реалізації алгоритмів.

Використання системи дає можливість відфільтрувати дані, здійснити форматування та певну корекцію. Робочі дані будуть внесені до бази даних для подальшого використання для подальшого використання. Реалізовані модулі для різноманітного представлення даних, надають можливість проаналізувати різноманітні показники, відслідкувати статистику, а також розрахувати можливість зміни певних показників у майбутньому (передбачення результатів звітності наступного року). Використання постреляційної бази даних забезпечує оперативний пошук даних, попередню селекцію та упорядкування.

У документі викладено основні етапи розробки, модулі для роботи даного проекту, структура об'єктів бази даних та опис даного програмного забезпечення.

Зміст

Вступ

Сфера інформаційних технологій щороку розширює рамки діяльності, що в свою чергу утворює потребу спеціалістів на ринку праці. Охоплення великої області застосувань стає причиною розробки та популяризації різноманітних технологій розробки, шо вказує на потребу вузько профільованості спеціалістів.

Однак, подібний вектор розвитку спричиняє певну складність для осіб, які зацікавлень у зміні сфери діяльності на сферу IT. Стає складно знайти відповідну нішу, яка відповідатиме потребам працівника. Таким чином з’являється потреба доступного аналізу ринку, порівняння умов праці, заробітної плати, можливості професійного росту та іншого.

Для отримання інформації в презентаційному форматі і буде реалізований програмний продукт, що проаналізує дані, та відобразить результати аналізу.

Технології використані при виконанні курсового проекту

Проект виконаний мовою програмування Python 3. Такий вибір зроблений через ряд причин:

- Динамічність, що проявляється у можливості легко перевизначати змінні, створювати гнучкі структури даних та спростити маніпуляції над ними
- Велика база сторонніх бібліотек різноманітної сфери застосування. Популярність мови серед спеціалістів наукової сфери, є причиною створення математичних та наукових бібліотек, що значно спрощують реалізацію програмних систем, що базуються на комплексному теоретичному підгрунті
- Значна перевага з точки зору аналізу даних перед іншими мовами програмування. Це пов’язано з специфічною оптимізацією мови, саме для роботи з великими об’ємами даних

Використані бібліотеки

pymongo - робота з базою даних MongoDB

ploty - створення графіків та діаграм у форматі html сторінок

pandas - робота з дата фреймами

scipy - аналіз даних, розрахунок лінійної регресії

База даних

Для розробки даної курсової роботи було обрано базу даних MongoDB. Це постреляційна документо-орієнтована база даних. Подібна модель дозволяє легко маніпулювати будь-якими структурами даних. MongoDB є швидкою та масштабованою системою, що надає простий доступ до даних шляхом запитів, які мають гнучкий функціонал для отримання потрібних документів.

MongoDB була обрана оскільки база даних міститиме ненормалізовані дані (результати опитувань працівників IT індустрії), зручні валідаційні схеми дозволяють встановити необхідний контроль даних, а наявність різноманітних бібліотек спрощують програмну реалізацію взаємодії з базою даних.

Середовище розробки

Розробка курсової здійснювалося в IDE Pycharm Educational Edition. Середовище забезпечує коректне відстежування синтаксису та імпорт усіх необхідних бібліотек.

Структура Бази даних

База даних містить одну колекцію - колекцію звітів працівників сфери IT.

db-coursework-data:

_id: ObjectId

city: string - місто де працює працівник

salary: number - місячна зарплата працівника

promotion: number - зміна заробітної плати протягом року

position: string - посада

language: string - мова програмування

year: year - рік звітування

month: month - місяць звітності

Для бази даних було реалізовано реплікацію, було створено 3 сервери (рис. 111), що утворюють собою структуру - сервера арбітра, головного сервера та другорядного сервера.

Схема взаємодії серверів має наступний вигляд

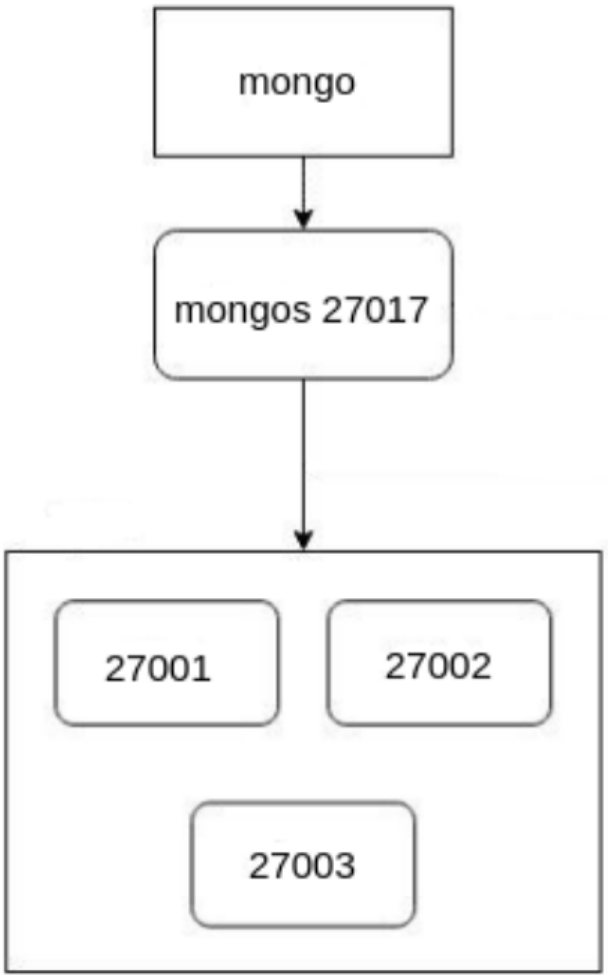

рис 1. Схема взаємодії серверів.

Опис програмного забезпечення

Загальна структура програмного забезпечення

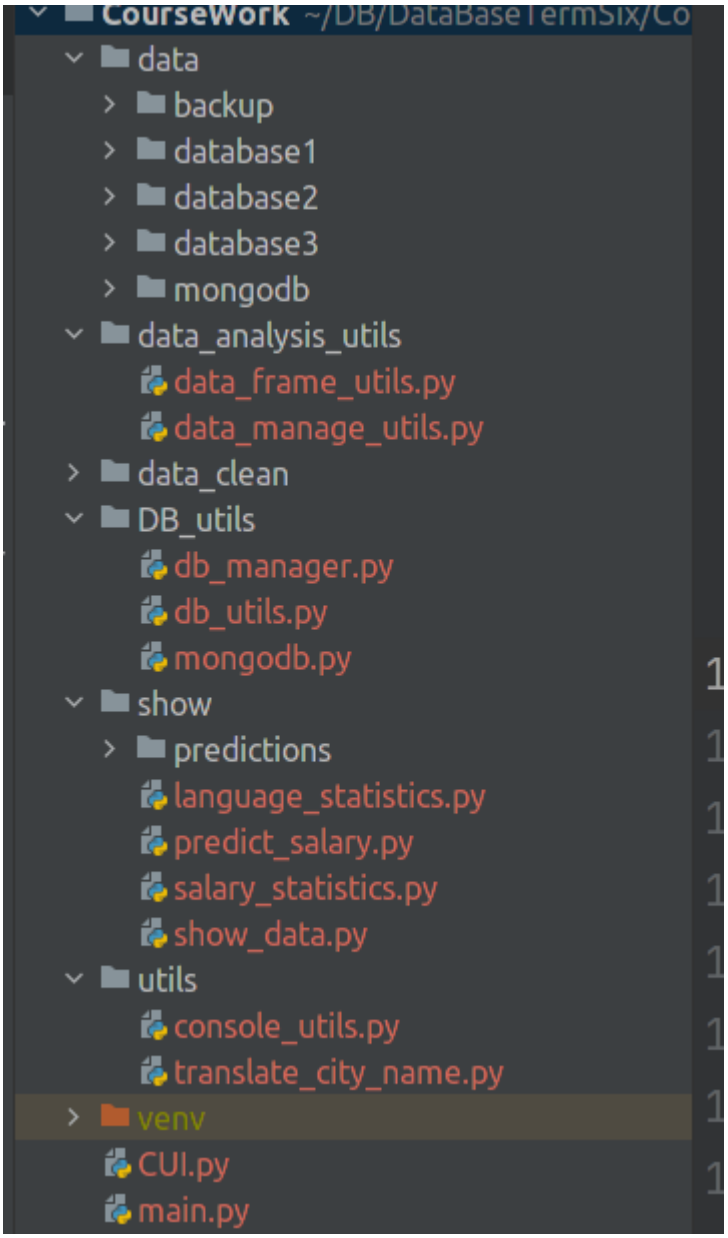

рис. 2, структура модулів

Модуль data:

Директорія де зберігаються бази даних реплікації та дані для відновлення (бекап)

Директорія data_analysis_utils:

Директорія де зберігаються модулі для роботи з дата фреймами:

manage_utils - опрацьовує вхідний дата сет, корегує та форматує вхідні дані, повертає відповідний дата фрейм

frame_utils - забезпечує систему необхідним функціоналом для отримання даних з дата фрейму

Директорія DB_utils:

Директорія містить модулі для взаємодії з базою даних:

db_manager: модуль що представляє сутність базт даних і містить набір необхідних модулів

db_utils: модуль, що містить додатковий функціонал для роботи основного модуля бази даних

mongodb: модуль конфігурації бази даних

Директорія show:

Директорія містить модулі представлення функціоналу системи:

language_statistics: модуль генерації діаграми популярності мов програмування в певний рік

predict_salary: модуль розрахунку передбачення заробітної плати на наступний рік на посаді для конкретної мови програмування

salary_statistics: модуль генерації графіків залежності заробітної плати, посади та року

show_data: модуль отримання необхідних даних для роботи інтерфейсу

Модуль CUI:

Модуль консольного інтерфейсу для зручнішого використання системи аналізу

Опис алгоритму роботи програми

При першому запуску (коли колекція бази даних порожня), модуль для створення дата фрейму відкриває список файлів - звітів працівників в форматі .csv (ресурсом в рамках виконання курсової роботи є сайт dou.ua), після виконання обробки даних, формується дата фрейм з усіма необхідними даними у формалізованому вигляді.Після цього дані заносяться до бази даних.

З моменту внесення даних до БД, програма готова до генерації графіків та діаграм.

Для подальшої роботи необхідно обрати відповідний пункт в меню програми (рис. 3).

Меню демонстрації статистики для обраної мови програмування, запускає модуль, який в свою чергу виконує запити до БД на отримання необхідних, після чого відбувається виконання математичних операцій результат яких демонструє графік (приклад на рис. 4)

Меню демонстрації популярності мов програмування, отримує дані по конкретно вказаному року, після чого дані представляються в груговій діаграмі (приклад рис. 5)

Меню передбачення заробітної плати на наступний рік за попередньо обраною мовою програмування отримує необхідні дані з БД, після чого виконуються математичні обчислення, розрахунок лінійної регресій, на основі якої визначається середню заробітня плата для різних посад, далі дані зберігаються у файлі формату .csv (приклад на рис. 6)

Висновок

Завдяки даній курсовій роботі виконавцем (студентом) були набуті навички взаємодії з великими об’ємами даних з використанням постреляційних баз даних, оформлення проектної документації у вигляді текстового та графічного матеріалу. Було розроблено та сконфігуровано програмне забезпечення для взаємодії та обробки великих масивів даних за допомогою постреляційної бази даних з використанням технік забезпечення відмовостійкості та доступності серверів.

Результат роботи у вигляді програмного коду можна переглянути у Додатку.

Література

1. Pymongo Tutorial [електронний ресурс] - https://pymongo.readthedocs.io/en/stable/tutorial.html

2. Репликация в MongoDB [Електронний ресурс]. – https://linux-notes.org/replikatsiya-v-mongodb/

3. Python: [Електронний ресурс]. – https://www.python.org/

4. Replication in MongoDB: [Електронний ресурс]. – https://docs.mongodb.com/manual/replication/

5. Pandas Documentation [Електронний ресурс]. - https://pandas.pydata.org/docs/

6. Python Machine Learning Linear Regression [Електронний ресурс] - https://www.w3schools.com/python/python_ml_linear_regression.asp

7. Ploty Python Open Source Graphic Library [Електронний ресурс] - https://plotly.com/python/

Додатки

Посилання на код розробленого програмного застосунку:
https://github.com/Bodichelly/DataBaseTermSix/tree/master/CourseWork

```
#################### Select Command ####################
[0] Salary statistics
[1] Languages popularity statistics per year
[2] Predict salary for a specific language
[3] DB manipulation
[4] Exit
Enter code:
```

рис. 3 Головне меню програмного застосунку

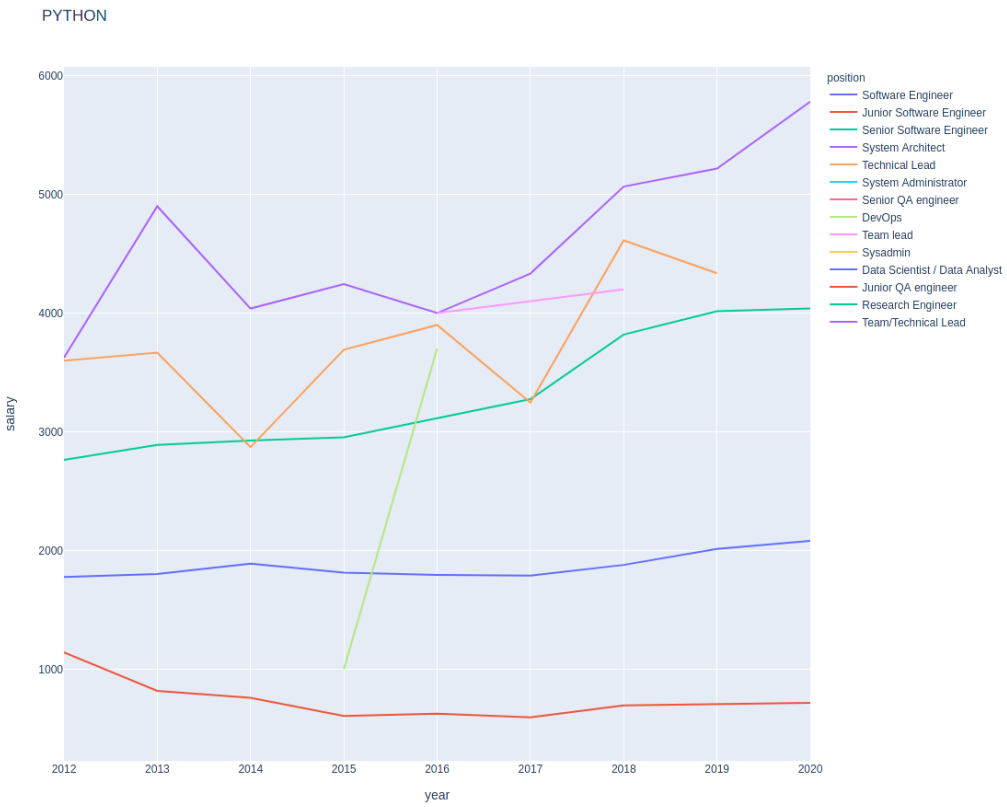

рис.4.1 Статистика заробітних плат для мови Python

рис.4.2 Статистика заробітних плат для мови Java

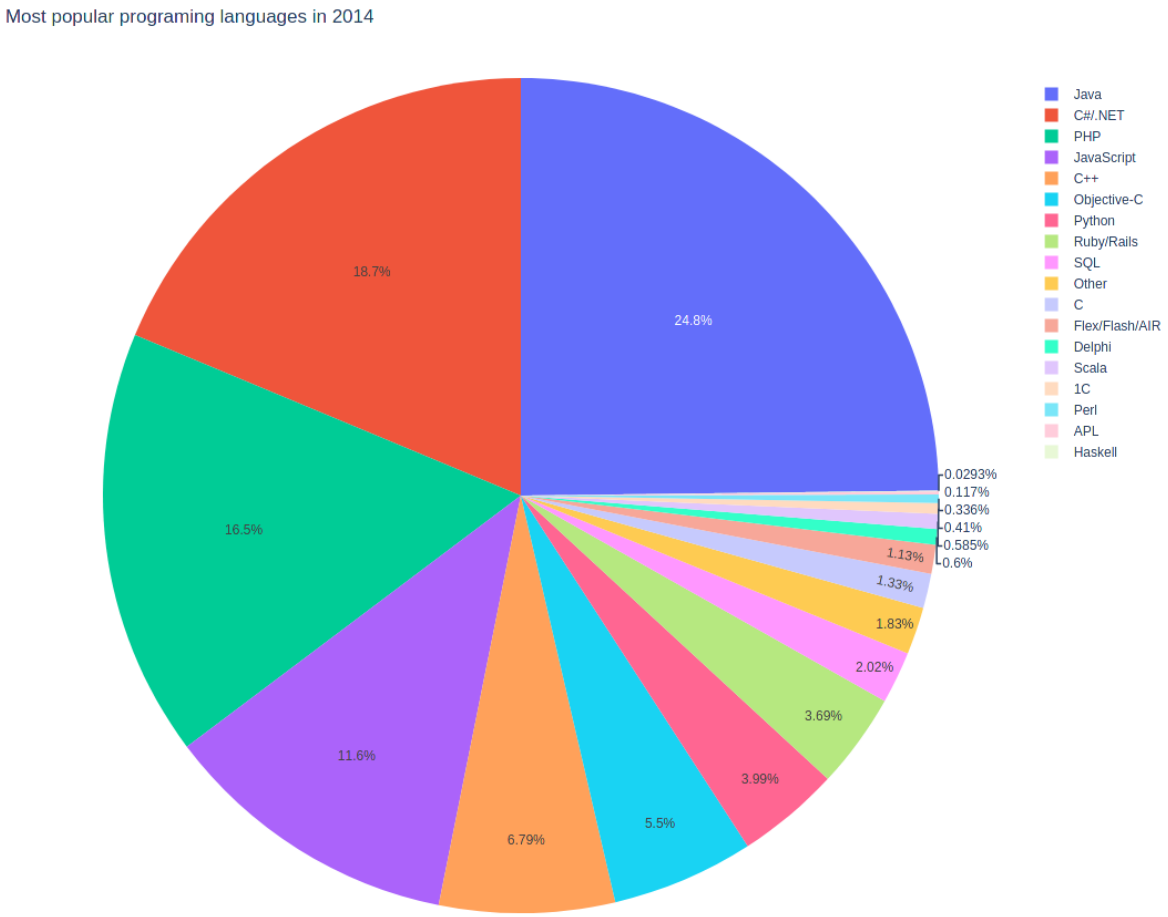

рис.5.1 Діаграма популярності мов програмування станом на 2014 рік

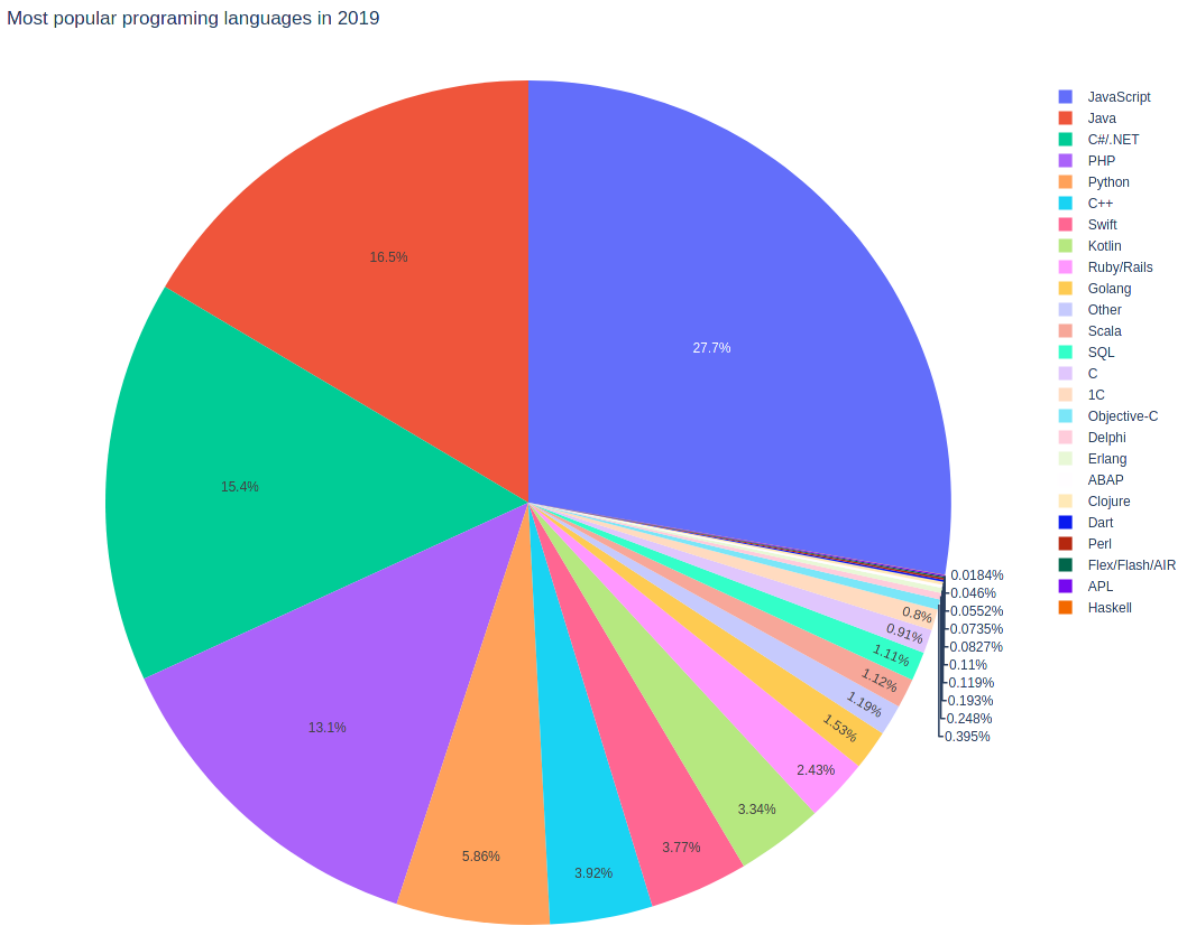

рис.5.2 Діаграма популярності мов програмування станом на 2019 рік

JavaScript_prediction.csv

```
position,salary
Junior Software Engineer,710
Software Engineer,1767
Technical Lead,4140
Senior Software Engineer,3193
System Architect,4994
Junior QA engineer,800
Support,700
Sales manager,625
Senior Project Manager / Program Manager,2000
HTML Coder,1225
QA engineer,1200
Team lead,3000
Scrum Master,3000
Project manager,1000
Team/Technical Lead,4196
```

рис.6.1 Передбачення середньої заробітної плати для спеціалістів мови JavaScript

Python_prediction.csv

```
position,salary
Software Engineer,1888
Junior Software Engineer,791
Senior Software Engineer,3269
System Architect,4723
Technical Lead,4258
System Administrator,950
Senior QA engineer,2500
DevOps,3700
Team lead,4200
Sysadmin,3000
Data Scientist / Data Analyst,750
Junior QA engineer,520
Research Engineer,1500
Team/Technical Lead,4102
```

рис.6.2 Передбачення середньої заробітної плати для спеціалістів мови Python