УДК 004.4 https://doi.org/10.35546/kntu2078-4481.2023.1.20

І. О. КАНДИБА
PhD, старший викладач кафедри інженерії програмного забезпечення
Чорноморський національний університет імені Петра Могили
ORCID: 0000-0002-8589-4028

Г. В. ГОРБАНЬ
кандидат технічних наук, доцент,
доцент кафедри інженерії програмного забезпечення
Чорноморський національний університет імені Петра Могили
ORCID: 0000-0002-6512-3576

Н. В. ГОНЧАРОВА
аспірант кафедри інженерії програмного забезпечення
Чорноморський національний університет імені Петра Могили
ORCID: 0000-0001-5536-6200

Д. С. ГОНЧАРОВ
аспірант кафедри комп'ютерної інженерії
Чорноморський національний університет імені Петра Могили
ORCID: 0009-0004-1200-6677

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ ПРО ВИКОРИСТАННЯ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ КОМП'ЮТЕРІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ ЗАСОБАМИ PYTHON

В статті представлено дослідження інструментарію для реалізації первинного статистичного аналізу даних про використання обчислювальних ресурсів мережі закладу вищої освіти (ЗВО). Досліджено сучасні роботи, присвячені аналізу даних про використання обчислювальних ресурсів, та шляхи впровадження результатів цих досліджень. Окрім того, описано специфіку моніторингу використання обчислювальних ресурсів ЗВО. Наведено опис особливостей споживання обчислювальних ресурсів мережі шкідливим ПЗ. Досліджено можливості застосування інструментарію мови Python для аналізу апаратно-технічного стану мережі ЗВО та споживання обчислювальних ресурсів. Запропоновано архітектуру бази даних (БД) для зберігання отриманих в результаті моніторингу даних про використання обчислювальних ресурсів і наведено опис полів, де міститиметься основна інформація: використання центрального та графічного процесорів, завантаженість оперативної пам'яті, заповнення накопичувача даних. Більш того, проаналізовано можливості бібліотеки Pandas в контексті первинного статистичного аналізу даних про використання обчислювальних ресурсів. Таким чином, було визначено особливості структури для зберігання даних DataFrame. Описано методи завантаження даних з БД до структури DataFrame. Розглянуто засоби динамічного відображення даних з допомогою Jupyter Notebook. Представлено метод реалізації первинного статистичного аналізу, а саме розрахунку: мінімального, максимального, середнього значень, квартилів, моди і медіани. Наведено опис використання бібліотек Matplotlib та Seaborn для візуалізації отриманих результатів. Розглянуто можливість використання гістограм для порівняння результатів моніторингу за кілька різних днів. Розглянуто можливість побудови діаграми розсіювання на основі отриманих даних про використання обчислювальних ресурсів. Виділено основні переваги застосування розробленого ПЗ: можливість визначення взаємозв'язків показників, діагностування наявності шкідливого ПЗ та прогнозування необхідних обчислювальних ресурсів для коректної роботи мережі ЗВО. Визначено подальші шляхи розвитку запропонованого ПЗ статистичного аналізу даних про використання обчислювальних ресурсів комп'ютерів ЗВО.

Ключові слова: *статистичний аналіз, Python, Pandas, Matplotlib, Seaborn.*

I. O. KANDYBA
PhD, Senior Lecturer at the Department of Software Engineering
Petro Mohyla Black Sea National University
ORCID: 0000-0002-8589-4028

H. V. HORBAN
Candidate of Technical Sciences, Associate Professor,
Associate Professor at the Department of Software Engineering
Petro Mohyla Black Sea National University
ORCID: 0000-0002-6512-3576

N. V. HONCHAROVA
Postgraduate Student at the Department of Software Engineering
Petro Mohyla Black Sea National University
ORCID: 0000-0001-5536-6200

D. S. HONCHAROV
Postgraduate Student at the Department of Computer Engineering
Petro Mohyla Black Sea National University
ORCID: 0009-0004-1200-6677

STATISTICAL DATA ANALYSIS ON THE COMPUTING RESOURCES USE OF UNIVERSITY COMPUTERS BY MEANS OF PYTHON

The article presents a study of tools for implementing primary statistical data analysis on the use of university network computing resources. Modern researches on the data analysis on the computing resources use and ways to implement results of them are studied in it. Moreover, peculiarities of monitoring the use of higher education institutions computing resources are described. Also, features of malware consumption of computing resources are described in this research. The article discusses a possibility of using Python tools for analysing the hardware and technical state of a university network and the computing resources consumption. Thus, it proposes an architecture of a database for storing the obtained data on the use of computing resources. In addition, a description of fields that store the main data is given: the use of the central and graphics processors, the load of RAM, and the data storage device filling. The authors analysis possibilities of the Pandas library in a context of primary statistical data analysis on the use of computing resources. Moreover, features of the DataFrame data storage structure are investigated in this research. Therefore, article describes methods of loading data from the database into the DataFrame structure and means of dynamic data display with Jupyter Notebook. Also, the authors consider a method of implementing primary statistical analysis, namely the calculation of: minimum, maximum, average, quartiles, mode, median. In addition, a description of the use of the Matplotlib and Seaborn libraries for visualizing the results is given. The publication considers possibility of using histograms to compare monitoring results for several different days and building a scatter plot based on the obtained data on the use of computing resources. The main advantages of using the developed software are highlighted: the ability to determine the interrelationships of indicators, diagnose the presence of malware and predict the necessary computing resources for the correct operation of the university network. Further ways of development of the proposed software for statistical data analysis on the use of computing resources of higher education institutions` computers are determined.

Key words: *statistical analysis, Python, Pandas, Matplotlib, Seaborn.*

Постановка проблеми

Аналіз даних про використання обчислювальних ресурсів комп'ютерів поєднаних певною локальною мережею застосовується у різних випадках, наприклад: контролю мережевого трафіка [1, с. 2], розподілу ресурсів та керування віртуальними середовищами [2, с. 41] тощо. Моніторинг та аналіз даних про використання обчислювальних ресурсів комп'ютерів закладу вищої освіти може допомогти оптимізувати навчальний процес [3, с. 44]. Проблема полягає у відсутності кросплатформового засобу для вирішення цієї задачі.

Аналіз останніх досліджень і публікацій

Сучасні дослідження в галузі аналізу даних про використання обчислювальних ресурсів мають різне спрямування. В роботі [1, с. 2] наведено опис методу використання результатів статистичного аналізу даних моніторингу та контролю навантаження програмно-конфігурованої мережі. В роботі наведено детальний опис архітектури системи для контролю мережевого трафіка на основі аналізу статистичних даних, але автори приділяють недостатньо уваги засобам аналізу статичних даних.

Робота [2, с. 41] присвячена оптимізації розподілу обчислювальних ресурсів у віртуальних середовищах. Наведено опис методу планування необхідного обсягу обчислювальних ресурсів на основі результатів статистичного аналізу даних моніторингу віртуальних середовищ. Авторами запропоновано використання засобів моніторингу та аналізу використання обчислювальних ресурсів, що вбудовані у засоби віртуалізації. Однак, у роботі не розглянуто можливість використання кросплатформових засобів моніторингу обчислювальних ресурсів з подальшим застосуванням статистичного аналізу.

В закладах вищої освіти аналіз даних використовується для вдосконалення навчального процесу. Робота [4, с. 35] присвячена автоматизованому збору та статистичному аналізу даних про відвідування занять. Описано архітектуру та технології системи контролю та аналізу відвідуваності закладу освіти. Представлено алгоритм реалізації регресивного аналізу даних щодо відвідування занять, але автори приділяють недостатньо уваги засобам аналізу даних про використання обчислювальних ресурсів.

Виконаний аналіз демонструє, що моніторинг даних про використання обчислювальних ресурсів є актуальним напрямом досліджень, але питання статистичного аналізу цих даних потребують подальших досліджень. Зокрема недостатньо уваги приділено методам первинного статистичного аналізу та засобам реалізації їх.

Формулювання мети дослідження

Метою дослідження є створення інформаційної системи аналізу даних про використання обчислювальних ресурсів закладу вищої освіти (ЗВО). Створення відповідної інформаційної системи дозволить оптимізувати розподіл обчислювальних ресурсів ЗВО, виявити несправності та визначити наявність шкідливого програмного забезпечення (ПЗ) в мережі.

Основні задачі, які вирішуються для досягнення заявленої мети:

1) провести дослідження впливу шкідливого ПЗ на інтенсивність використання обчислювальних ресурсів;
2) проаналізувати засоби моніторингу використання обчислювальних ресурсів ЗВО;
3) розробити ПЗ реалізації первинного статистичного аналізу та візуалізації даних.

Викладення основного матеріалу дослідження

Навчання фахівців у галузі інформаційних технологій (ІТ) вимагає наявності різноманітного програмного та апаратного забезпечення. Цей факт обумовлює складність побудови локальної мережі ЗВО та моніторингу споживання обчислювальних ресурсів.

Моніторинг обчислювальних ресурсів ЗВО корисний виключно за наявності засобів його аналізу. Наприклад, первинний статистичний аналіз дозволяє побудувати варіаційний ряд (гістограми) на основі ранжування результатів проведених вимірів споживання обчислювальних ресурсів. Симетрична відносно центру гістограма свідчить про відсутність проблем з обчислювальними ресурсами у мережі, а у випадку наявної асиметрії можна констатувати певні проблеми у мережі ЗВО.

Особливо важливим є виявлення шкідливого ПЗ, що не завжди можливо вчасно виконати спеціалізованими програмними засобами та запобігти нанесенню значної шкоди. Прикладом розповсюдження шкідливого ПЗ є Petya (вірус шифрувальник), що у 2017 році став причиною масштабних втрат інформації та спричинив багато проблем у різних галузях, пов'язаних з ІТ [5, с. 122]. Це шкідливе ПЗ шифрує файли на жорсткому диску та надсилає їх зловмиснику. Цей процес призводить до споживання великої кількості обчислювальних ресурсів: процесорного часу, оперативної пам'яті, використання накопичувачів даних [5, с. 122]. У цьому контексті задача моніторингу та аналізу даних про використання обчислювальних ресурсів з метою виявлення шкідливого ПЗ постає особливо гостро.

Особливостями мережі ЗВО є те, що лабораторії, які є частиною навчального закладу, складаються з програмного та апаратного забезпечення різного типу. Різні операційні системи та архітектури комп'ютерів потребують окремих спеціалізованих засобів моніторингу апаратно-технічного стану та використання обчислювальних ресурсів [3, с. 44].

Використання можливостей мови програмування Python дозволяє врахувати зазначені особливості та реалізувати збір даних про використання апаратного забезпечення мережі ЗВО [3, с. 44]. Таким чином, зберігання зібраних даних вимагає наявності спеціалізованого сховища системи керування базами даних (СКБД) SQLite.

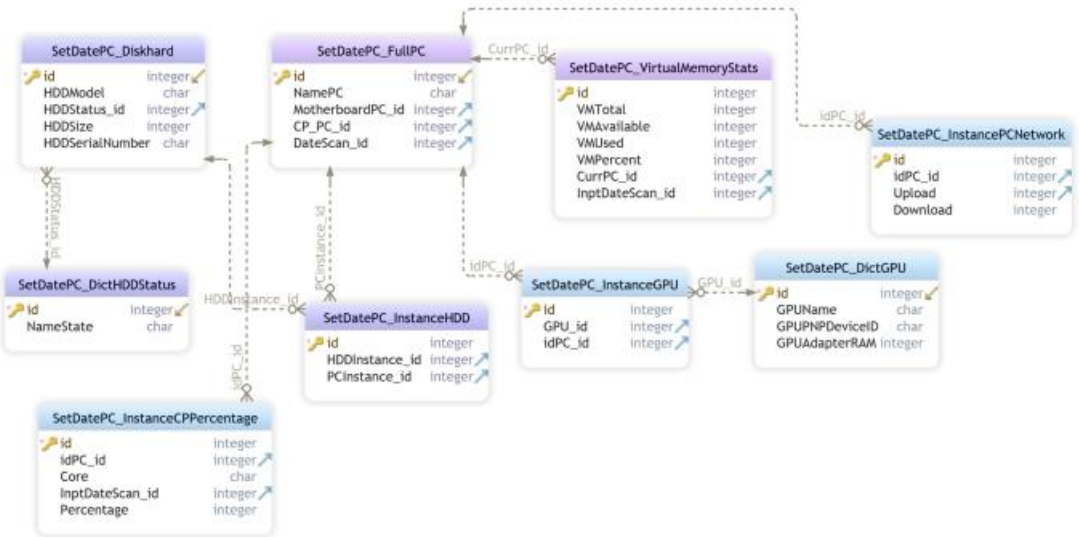


Рис. 1. Фрагмент даталогічної моделі БД для зберігання даних моніторингу використання обчислювальних ресурсів

В наведеному фрагменті БД подальшого аналізу вимагають поля:

– **CP_Petcentage** в таблиці **SetDataPC_InstanceCPPrecentage** зберігає дані про використання процесорного часу;

– **VMUsed** в таблиці **SetDataPC_VirtualMemoryStats** зберігає дані про використання оперативної пам'яті та файлу підвантаження;

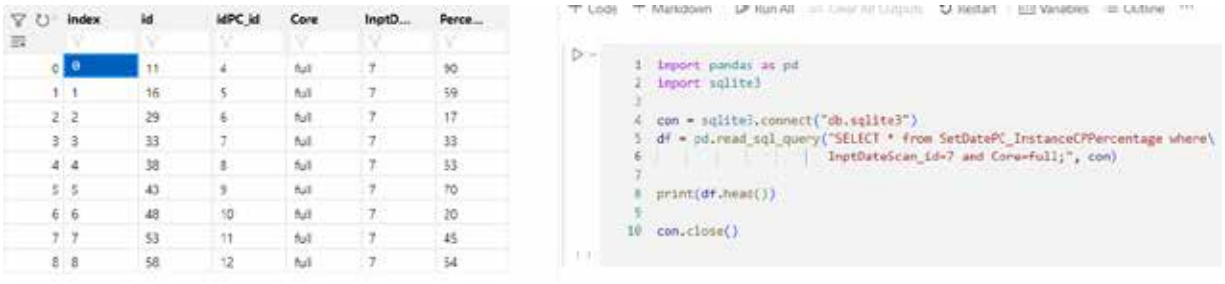
– **GPU_Petcentage** в таблиці **SetDataPC_InstanceGPU** зберігає дані використання графічного процесора;

– **SizeUsedPercent** в таблиці **SetDataPC_PCPartition** зберігає дані про використаний об'єм накопичувача даних;

– **Upload** і **Download** в таблиці **SetDataPC_InstancePCNetwork** зберігає дані про об'єм надісланих та прийнятих мережею даних.

З отриманого набору вхідних даних проведено первинний статистичний аналіз за допомогою мови Python з використанням інструментарію Pandas [6, с. 116]. Ця бібліотека спрощує завантаження даних і реалізує методи обробки та аналізу даних. Pandas реалізує дві структури для зберігання даних: Series – одномірний індексований масив даних фіксованого типу та DataFrame – двовимірна структура даних, кожен стовпець якої містить дані одного типу. DataFrame доцільно застосовувати для представлення збереження даних: рядки відповідають описам ознак окремих об'єктів, а стовпці відповідають ознакам.

Для реалізації первинного статистичного аналізу даних про використання обчислювальних ресурсів ЗВО результат моніторингу завантажено у DataFrame. Для цього використано метод read_sql_query, що забезпечує виконання запиту до БД з підтримкою SQL. Додатково використано Jupyter Notebook, що дозволяє динамічне відображення результатів статистичної обробки. На рис. 2 представлено результат імпортування даних в DataFrame.



	index	id	idPC_id	Core	InptD...	Perce...
0	0	11	4	full	7	90
1	1	16	5	full	7	59
2	2	29	6	full	7	17
3	3	33	7	full	7	33
4	4	38	8	full	7	53
5	5	43	9	full	7	70
6	6	48	10	full	7	20
7	7	53	11	full	7	45
8	8	58	12	full	7	54

Рис. 2. Результат імпорту даних з SQL до DataFrame

Відображення DataFrame несе ілюстративну функцію, але для статистичного аналізу необхідно використання методу describe( ) [7 с. 108]. Цей метод виконує первинний статистичний аналіз та розраховує наступні характеристики: кількість ненульових елементів (count); мінімальне (min), максимальне (max), середнє значення (mean); перший (75%), другий(50%) та третій (75%) квартилі.

Отже, виконавши SQL запит «SELECT * from SetDatePC_InstanceCPPercentage», можна отримати дані про навантаження центрального процесора (ЦП), але для початку аналізу необхідно відібрати виміри проведені одночасно та для всього ЦП в цілому, а не для певного ядра (рис. 3).

```
[25]  ✓ 0.1s

count     9.000000
mean     49.000000
std      23.452079
min      17.000000
25%      33.000000
50%      53.000000
75%      59.000000
max      90.000000
Name: Percentage, dtype: float64
```

Рис. 3. Результат розрахунку дескриптивної статистики

Первинний статистичний аналіз має включати в себе також визначення моди та медіани. Мода – це значення, що зустрічається в множині найчастіше. Однією з ознак наявності шкідливого ПЗ в мережі є величина моди

ознаки навантаження ЦП у розмірі 100. В бібліотеці Pandas пошук моди реалізується за допомогою методу mode( ). Медіана – це величина ознаки, що розташована посередині ранжованого ряду ознак. Можливо використати метод mean( ) для визначення середнього значення використання певного обчислювального ресурсу.

```
[9]
import pandas as pd
import sqlite3

con = sqlite3.connect("db.sqlite3")
df = pd.read_sql_query("SELECT * from SetDatePC_InstanceCPPercentage;", con)
mean_1=df['Percentage'].mean()
mode_1=df['Percentage'].mode()
print(f'Mean {mean_1}')
print(f'Mode {mode_1}')
con.close()
```

✓ 0.0s

```
Mean 48.888888888888886
Mode 0    53
```

Рис. 4. Визначення моди та медіани використання ЦП

Після визначення всіх перерахованих вище характеристик отримані дані моніторингу обчислювальних ресурсів можливо візуалізувати. Доцільно застосовувати бібліотеки Matplotlib та Seaborn для відображення гістограм, діаграми розсіювання ознак, кореляційних матриць тощо [8, с. 151].

Гістограма являє собою представлення розподілу числових ознак визначених характеристик. Як вже згадувалось вище, наявність асиметрії в гістограмі даних про використання обчислювальних ресурсів ЗВО свідчить про проблеми в мережі. Діагностування проблем, таким чином, потребує відображення одночасно чотирьох характеристик: відсоток використання ЦП, відсоток використання графічного процесора, відсоток використання оперативної пам’яті, відсоток зайнятого об’єму накопичувача даних (рис. 5).

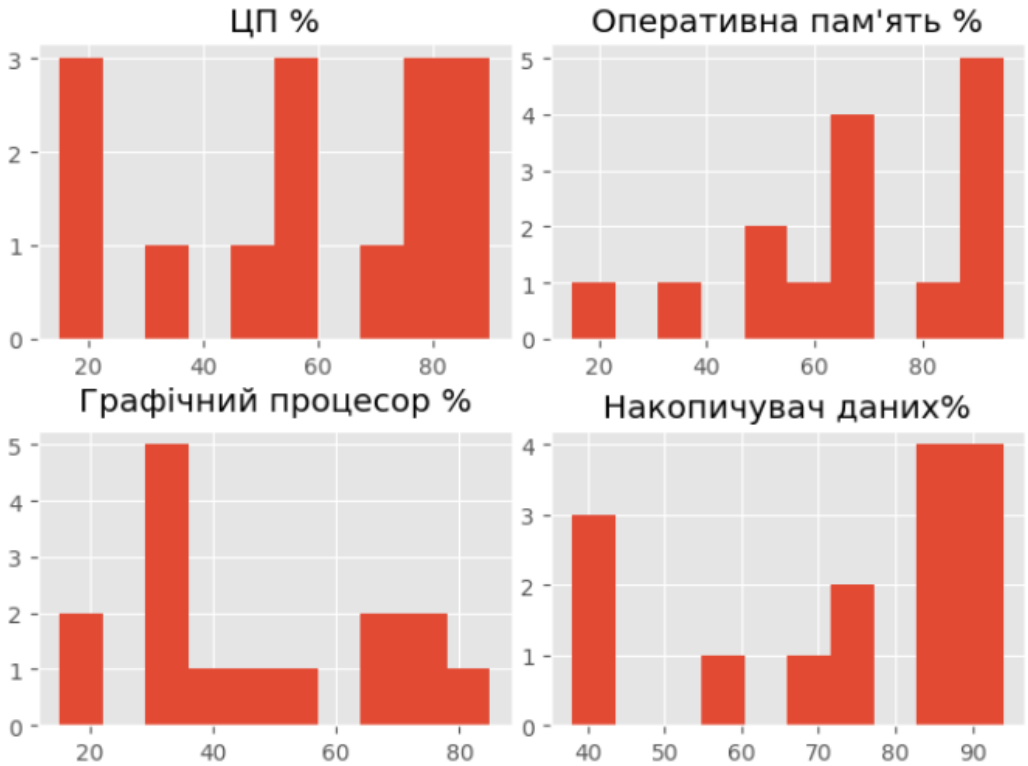


Рис. 5. Гістограми використання обчислювальних ресурсів

Гістограми на рисунку 5 демонструють, що найбільше, в одній з лабораторій мережі ЗВО, використовується оперативна пам’ять, але більшість ЦП навантажена не більше, ніж на 50%, що може вказувати на відсутність

шкідливого ПЗ в мережі. Проте, залишається потреба визначати деякі проблеми апаратного забезпечення, наприклад, падіння швидкості запису даних на накопичувач (HDD/SSD) призводить до збільшення навантаження на оперативну пам'ять. Визначити проблеми такого плану можна, використавши діаграму розсіювання. Діаграма розсіювання, також відома як scatter plot – це графік, який використовується для відображення залежності між двома змінними. Діаграма розсіювання складається з точок, розташованих на графіку, де кожна точка представляє значення двох змінних для одного спостереження. Горизонтальна вісь зазвичай відображає значення однієї змінної, а вертикальна вісь – значення іншої змінної. Інструментарій Matplotlib та Seaborn дозволяють виконувати побудову різних варіацій діаграм розсіювання. Наприклад, метод jointplot( ) дозволяє малювати комбінований графік гістограм та діаграм розсіювання.

Використовуючи зазначений інструментарій, можна дослідити залежності між ознаками попарно. Так, наприклад, можна побудувати діаграму розсіювання для відсотка завантаженості ЦП та використання оперативної пам'яті. (рис. 6).

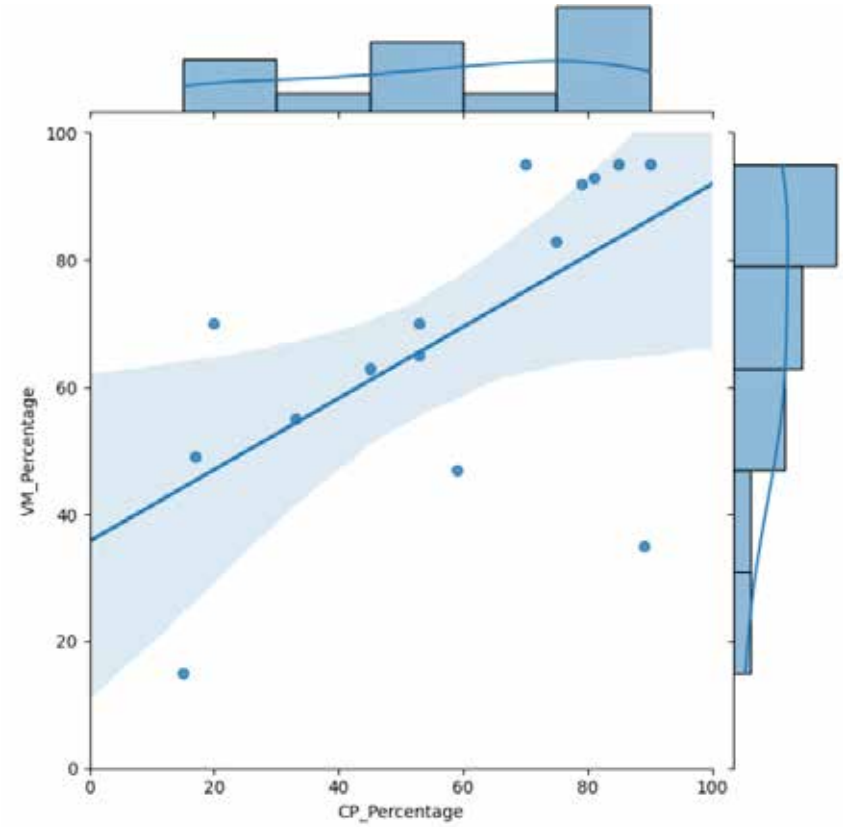


Рис. 6. Діаграма розсіювання даних про використання оперативної пам'яті та навантаження ЦП

Отримане графічне представлення дозволяє швидко оцінити, як розподілені дані, та визначити, чи є залежність між двома ознаками. Представлені гістограми можуть бути використані для порівняння розподілів даних. Наприклад, для порівняння результатів моніторингу використання обчислювальних ресурсів в різні дні.

Гістограми та діаграми розсіювання можуть допомогти в прогнозуванні майбутніх тенденцій. У випадку, коли діаграма розсіювання показує сильну залежність між двома змінними, можна передбачити необхідність оновлення певного апаратного забезпечення в мережі ЗВО або визначити наявність шкідливого ПЗ та запобігти втраті даних.

Висновки

1) Досліджено вплив шкідливого ПЗ на рівень використання обчислювальних ресурсів. Наведено опис особливостей споживання обчислювальних ресурсів одним з найбільш відомих вірусів-шифрувальників Petya.

2) Проаналізовано інструментарій моніторингу використання обчислювальних ресурсів ЗВО, наведено архітектуру БД для зберігання зібраних даних.

3) На основі Pandas розроблено ПЗ для проведення первинного статичного аналізу. Запропоновано використання DataFrame для визначення розрахунку дескриптивної статистики, визначення моди та медіани. Застосовано бібліотеки Matplotlib та Seaborn для візуалізації результатів моніторингу використання обчислювальних ресурсів ЗВО. Наведено приклад застосування діаграм розсіювання для визначення взаємозв'язку певних ознак.

Застосунок розроблений мовою загального призначення Python, що підтримує велику кількість різного інструментарію, в тому числі реалізацію методів штучного інтелекту, що можуть бути інтегровані для проведення подальшого аналізу даних про використання обчислювальних ресурсів ЗВО. Результати роботи розробленого ПЗ можуть бути використані різними ЗВО для визначення наявності проблем апаратного забезпечення та запобігання розповсюдження шкідливого ПЗ.

Список використаної літератури

1. Leonardi L., Bello L. Lo, Agliano S. Priority-based bandwidth management in virtualized software-defined networks. Electronics. Vol. 9, Issue 6. P. 1-6.

2. Somani G., Chaudhary S. Application Performance Isolation in Virtualization. 2009 IEEE International Conference on Cloud Computing(2009). DOI:10.1109/CLOUD.2009.78. P. 41–48.

3. Кандиба І. О., Горбань Г. В., Фісун М. Т., Ткаченко М. П. Дослідження апаратно-технічного стану локальної мережі закладу вищої освіти засобами мови Python. Вчені записки Таврійського національного університету імені В. І. Вернадського, 2022. № 3 т.33 С. 44 -49.

4. Чупілко Т., Чупілко О., Мормуль М. Проектування і програмна реалізація автоматизованої системи відвідуваності та аналізу даних в закладах освіти. Інформаційні технології та комп'ютерна інженерія. Випуск. 56, Том 1. 2023. С. 35–43.

5. Aidan J. S., Verma H. K., Awasthi L. K. Comprehensive survey on petya ransomware attack. 2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)(2017). P. 122–125.

6. Bantilan N. Pandera: Statistical data validation of pandas dataframes. Proceedings of the Python in Science Conference (SciPy)(2020). P. 116–124.

7. Teoh T. T., Rong Z. Python for Data Analysis. Artificial Intelligence with Python. Springer, 2022. P. 107–122.

8. Bisong E., Matplotlib and seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. 2019. P. 151–165.

9. Горбань Г. В., Кандиба І. О., Антіпова К. О., Кірей К. О. Первинний та візуальний аналіз даних спортивних результатів з академічного веслування засобами мови python з використанням бібліотек PANDAS, MATPLOTLIB ТА SEABORN. Таврійський науковий вісник. Серія: Технічні науки. Том 3. С. 27–37.

10. Несвіт М. І., Несвіт К. В. Розв'язання математичних задач сучасними мовами програмування та технологіями. Модернізація вищої освіти в україні та проблеми управління якістю підготовки фахівців у технічному університеті : зб. матеріалів університетської. наук.-метод. конф. ХНУБА. Харків. С. 120–123.

References

1. Leonardi L., Lo Bello, L., & Agliano, S. (2020). Priority-based bandwidth management in virtualized software-defined networks. Electronics, 9(6), 1009. DOI: DOI:https://doi.org/10.3390/electronics9061009

2. Somani G., & Chaudhary S. (2009). Application Performance Isolation in Virtualization. 2009 IEEE International Conference on Cloud Computing, pp. 41–48. DOI:https://doi.org/10.1109/CLOUD.2009.78

3. Kandyba I. O., Horban H. V., Fisun M. T., Tkachenko, M. P. (2022). Analysis of the hardware and technical state of the local network of a university's using the languagePython. Scientific notes of Taurida National V.I. Vernadsky University". Series: Technical Sciences, 33(3), pp. 44-49. DOI:https://doi.org/10.32838/2663-5941/2022.3/07

4. Chupilko T. A., Chupilko O. S., Mormul M. F. (2023) Design and software implementation of the automated attendance system and data analysis in educational institutions. University of Customs and Finance, Dnipro, pp. 35-43 DOI:https://doi.org/10.31649/1999-9941-2023-56-1-35-43

5. Aidan J. S., Verma H. K., & Awasthi L. K. (2017). Comprehensive survey on petya ransomware attack. 2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS), pp.122–125. DOI:https://doi.org/10.1109/ICNGCIS.2017.30

6. Bantilan N. (2020). Pandera: Statistical data validation of pandas dataframes. Proceedings of the Python in Science Conference (SciPy), pp. 116–124.

7. Teoh T. T., & Rong Z. (2022). Python for Data Analysis. In Artificial Intelligence with Python (pp. 107–122). Springer. DOI: https://doi.org/10.1007/978-981-16-8615-3

8. Bisong E.(2019). Matplotlib and seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, P. 151–165. DOI: https://doi.org/10.1007/978-1-4842-4470-8

9. Horban H. V., Kandyba I. O., Antipova K. O., Kirei K. O. (2022). Primary and visual analysis of rowing performance data by means of python using Pandas, Matplotlib and Seaborn libraries. Taurida Scientific Herald. pp. 27-37 DOI: https://doi.org/10.32851/tnv-tech.2022.3.3

10. Nesvit M. I., Nesvit K. V. (2020) Solving mathematical problems in modern programming languages and technologies. Modernization of higher education in ukraine and problems of quality management training of technical specialists universities. Kharkiv: KNUBA. pp 120–123.