Міністерство освіти і науки України
Національний технічний університет України „КПІ”
Факультет інформатики та обчислювальної техніки

Кафедра автоматизованих систем обробки
інформації та управління

ЗВІТ

з лабораторної роботи № 4
дисципліни
“ТЕХНОЛОГІЇ ПАРАЛЕЛЬНОГО ПРОГРАМУВАННЯ В УМОВАХ ВЕЛИКИХ ДАНИХ”
на тему:

„Big Data з використанням засобів Apache Spark”

Виконали студенти
- *ІП-01мн Семченко Андрій*
- *ІП-01мн Кошовець Євген*
- *ІТ-01мн Васюк Владислав*
- *ІТ-01мн Минзар Богдан*

(№ групи, прізвище, ім’я, по батькові)

Прийняв *доц. Жереб К. А.*
(прізвище, ім’я, по батькові)

Київ 2021

ЗМІСТ

1 ПОСТАНОВКА ЗАДАЧІ

Необхідно реалізувати вирішення обраної задачі з використанням технології Apache Spark. Можна запустити реалізацію локально, але якщо є можливість запустити на розподіленій системі – за це можна отримати додаткові бали. Порівняти реалізації лабораторних робіт No3 та No4. Результатом виконання даної лабораторної роботи є працююча програма, а також звіт про використані технології та можливості, з результатами вимірів.

Ідея полягає у тому, щоб шляхом аналізу вмісту публічного репозиторію знайти потенційні вразливості, що дозволяють втрутитись у роботу програмного забезпечення. Причому вразливості не тільки у самій реалізації програмного забезпечення, але і вразливості, спричинені недбалим обігом sensitive data, наприклад:

- Зберігання ключів доступу у файлах, що відстежуються VCS
- Зберігання ключів доступу прямо у тексті програмного забезпечення
- Зберігання бекапів у файлах, що відстежуються VCS

Зберігання sensitive data у файлах, що відстежуються системою контролю версій призводить до того, що будь хто може завантажити ці дані і використати для втручання у роботу програмного забезпечення, викрадення даних користувачів, тощо.

2 ВИКОРИСТАНІ БІБЛІОТЕКИ, ФРЕЙМВОРКИ

Розроблена програма використовує фреймворк PySpark та написана на мові програмування Python 3.7 у вигляді Jupyter ноутбука. Мотивація використовувати Python – вже є досвід роботи з PySpark, тож це значно спростить процес розробки та тестування коду. Для роботи з Github було використано бібліотеку GitPython [1].

3 ОПИС РОБОТИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

3.1Загальна концепція

Програмне забезпечення завантажує вміст публічного Github-репозиторію у тимчасову директорію. Далі, кожен файл даної директорії аналізується на предмет наявності в ньому певних патернів, що можуть свідчити про вразливість. Сам паттерн описується за допомогою регулярного виразу.

3.2Деталі роботи програмного забезпечення

Загальний принцип роботи:

1. Програма запускається у вигляді Python ноутбука з використанням Jupyter. Налаштування підключення до Spark кластеру задано в коді програми та у файлі spark-defaults.conf.
2. Після створення контексту Spark, програма створює DataFrame зі списком посилань на git-репозиторії, що був попередньо заданий в коді програми (однак на вхід можна використати інші джерела).
3. За допомогою UDF [2] виконується завантаження git-репозиторіїв в тимчасові папки кожного з воркерів, після чого зміст та шлях кожного файла в репозиторії зберігається в DataFrame. На виході маємо таблицю з полями: repo (шлях до репозиторію), path (відносний шлях до файлу в репозиторії), content (зміст файлу якщо він текстовий).
4. DataFrame з попереднього кроку за допомогою ще однієї UDF аналізуємо на вразливості: перевіряється шлях до кожного файлу та вміст. На виході отримуємо DataFrame з полями: repo (шлях до репозиторію), path (відносний шлях до файлу в репозиторії), issueType (тип вразливості), issueDescription (опис вразливості), lineNumber (номер рядку в файлі де знайдено вразливість).
5. Результати роботи програми експортується у форматі JSON за допомогою бібліотеки pandas.

Так виглядає список Completed Jobs в Spark UI після виконання:

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	toPandas at <timed exec>:1 toPandas at <timed exec>:1	2021/12/22 07:03:51	8 s	1/1 (1 skipped)	200/200 (192 skipped)
0	toPandas at <timed eval>:1 toPandas at <timed eval>:1	2021/12/22 07:03:03	48 s	2/2	392/392

4 ОТРИМАНІ РЕЗУЛЬТАТИ

- **Вимірювання часу роботи в залежності від параметрів**

Всі тести проводились на Kubernetes кластері з 5 машин (на базі процесора Intel(R) Xeon(R) E-2286G CPU @ 4.00GHz, 6 ядер, 12 потоків; 64 GB оперативної пам’яті на кожній)

В кожному тесті змінено два основні параметри: максимальна кількість воркерів (spark.executor.instances) та кількість використаних ядер на кожному воркері (spark.executor.cores).

spark.executor.instances	spark.executor.cores	**time (seconds)**
1	**1**	**56**
1	**2**	**36**
1	**4**	**20.2**
2	**1**	**33.2**
2	**2**	**19.34**
2	**4**	**18.4**
4	**1**	**23.4**
4	**2**	**18.02**
4	**4**	**17.7**
8	**1**	**20.5**

8	**2**	**16.6**
8	**4**	**16.6**

Як бачимо, найшвидше програма відпрацьовує з кількістю воркерів = 8 та кількістю ядер більше 2. Також можна помітити, що в програмі є «вузьке» місце – завантаження репозиторіїв; при подальших тестах на більшій кількісті воркерів стало зрозуміло, що час не скорочується, оскільки завантаження найбільшого репозиторію займає близько 15 секунд.

5 ВИСНОВОК

В рамках даної лабораторної роботи було розроблено програмне забезпечення на основі Apache Spark, що проводить аналіз Github-репозиторіїв на наявність в них типових вразливостей.

В рамках реалізованого алгоритму вдалося отримати пришвидшення в порівнянні з однопоточною реалізацією у 3.5 рази на конкретно взятому прикладі навантаження.

Було порівняно результати роботи програми з різними налаштуваннями паралелізму, найкращих результатів (16.6с) досягли при використанні 8 воркерів та 2+ ядер CPU на кожному воркері.

5 ПОСИЛАННЯ

[1] https://gitpython.readthedocs.io/en/stable/

[2] https://spark.apache.org/docs/3.1.1/api/python/reference/api/pyspark.sql.functions.udf.html

6 ДОДАТОК 1 - ЛОГ РОБОТИ ПРОГРАМИ

```
CPU times: user 70.7 ms, sys: 7.35 ms, total: 78.1 ms
Wall time: 13.6 s
```

[6]:

	repo	path	content
0	https://github.com/mikemelon/java-signin.git	java-signin/src/main/resources/config/G5_110_i...	# G5教学楼110机房\n# 第1排 左侧\n172.19.13.14=(1,3)\n1...
1	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-2/week-2...	package com.itpro.blog.controllers;public clas...
2	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2020-2021/@semester_2/w3-1/h...	None
3	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-2/week-7...	None
4	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2020-2021/@semester_2/w7-1/F...	package com.netit.database;\n\npublic enum Dat...
...	...	...	...
8098	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2020-2021/@semester_2/w19-2/...	package com.trelloclone.trelloclone.repositori...
8099	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-2/week-6...	<%@ page contentType="text/html;charset=UTF-8"...
8100	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2020-2021/@semester_2/w6-2/F...	None
8101	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2020-2021/@semester_1/week-1...	<?xml version="1.0" encoding="UTF-8"?>\n<class...
8102	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-1/week-9...	package tests.test;\n\nimport config.PieceColo...

8103 rows × 3 columns

```
[Stage 3:=====================================================>  (192 + 4) / 200]
CPU times: user 27.7 ms, sys: 1.52 ms, total: 29.2 ms
Wall time: 3.23 s
```

:

	repo	path	issueType	issueDescription	lineNumber
0	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-2/week-2...	Log file	Log files might contain information such as re...	NaN
1	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2020-2021/@semester_2/w8-1/s...	Log file	Log files might contain information such as re...	NaN
2	https://github.com/mikemelon/java-signin.git	java-signin/src/main/webapp/image/logo.jpg	Log file	Log files might contain information such as re...	NaN
3	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-2/week-6...	Log file	Log files might contain information such as re...	NaN
4	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-2/week-8...	Log file	Log files might contain information such as re...	NaN
...	...	...	...	...	...
458	https://github.com/mikemelon/java-signin.git	java-signin/src/main/java/cn/lynu/lyq/signin/a...	AWS key	Potential AWS Access Key ID expose	21.0
459	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-1/week-1...	AWS key	Potential AWS Access Key expose	41.0
460	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-1/week-1...	AWS key	Potential AWS Access Key expose	94.0
461	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-1/week-1...	AWS key	Potential AWS Access Key expose	105.0
462	https://github.com/mihail-petrov/netit-webdev-...	netit-webdev-java/2019-2020/@semester-1/week-1...	AWS key	Potential AWS Access Key expose	119.0

463 rows × 5 columns

```
▾ 0:
    repo: "https://github.com/mihail-petrov/netit-webdev-java.git"
    path: "netit-webdev-java/2019-2020/@semester-2/week-21-1/blog/blog/src/main/java/com/itpro/blog/controllers/BlogController.java"
    issueType: "Log file"
    issueDescription: "Log files might contain information such as references to secret HTTP endpoints, session IDs, user information, passwords and API keys."
    lineNumber: null
▾ 1:
    repo: "https://github.com/mihail-petrov/netit-webdev-java.git"
    path: "netit-webdev-java/2020-2021/@semester_2/w8-1/src/main/java/com/netit/logic/PageViewAggregator.java"
    issueType: "Log file"
    issueDescription: "Log files might contain information such as references to secret HTTP endpoints, session IDs, user information, passwords and API keys."
    lineNumber: null
▾ 2:
    repo: "https://github.com/mikemelon/java-signin.git"
    path: "java-signin/src/main/webapp/image/logo.jpg"
    issueType: "Log file"
    issueDescription: "Log files might contain information such as references to secret HTTP endpoints, session IDs, user information, passwords and API keys."
    lineNumber: null
▾ 3:
    repo: "https://github.com/mihail-petrov/netit-webdev-java.git"
    path: "netit-webdev-java/2019-2020/@semester-2/week-6-1/HelloWebWorld/web/login.jsp"
    issueType: "Log file"
    issueDescription: "Log files might contain information such as references to secret HTTP endpoints, session IDs, user information, passwords and API keys."
    lineNumber: null
▾ 4:
    repo: "https://github.com/mihail-petrov/netit-webdev-java.git"
    path: "netit-webdev-java/2019-2020/@semester-2/week-8-1/HelloWebWorld/web/login.jsp"
    issueType: "Log file"
    issueDescription: "Log files might contain information such as references to secret HTTP endpoints, session IDs, user information, passwords and API keys."
    lineNumber: null
▾ 5:
    repo: "https://github.com/mihail-petrov/netit-webdev-java.git"
    path: "netit-webdev-java/2019-2020/@semester-2/week-22-1/blog/blog/src/main/java/com/itpro/blog/models/request/HTTPRequest.java"
    issueType: "Log file"
    issueDescription: "Log files might contain information such as references to secret HTTP endpoints, session IDs, user information, passwords and API keys."
```