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 № 3
дисципліни
“ТЕХНОЛОГІЇ ПАРАЛЕЛЬНОГО ПРОГРАМУВАННЯ В УМОВАХ ВЕЛИКИХ ДАНИХ”
на тему:

„Big Data з використанням засобів Apache Hadoop”

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 підходу MapReduce та технології Apache Hadoop. Можна запустити реалізацію локально, але якщо є можливість запустити на розподіленій системі – за це можна отримати додаткові бали. 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

Додаток було розроблено мовою програмування Java. Для роботи з Github було використано бібліотеку JGit [1].

Для логування роботи програми була використана бібліотека Log4j2. Робота з JSON організована засобами бібліотеки Jackson Json.

Для спрощення читабельності коду застосовано бібліотеку Lombok, що автоматично генерує boilerplate код.

Код запущено на локальному кластері Hadoop 3.2.1, запущеному за допомогою docker compose [2].

3 ОПИС РОБОТИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

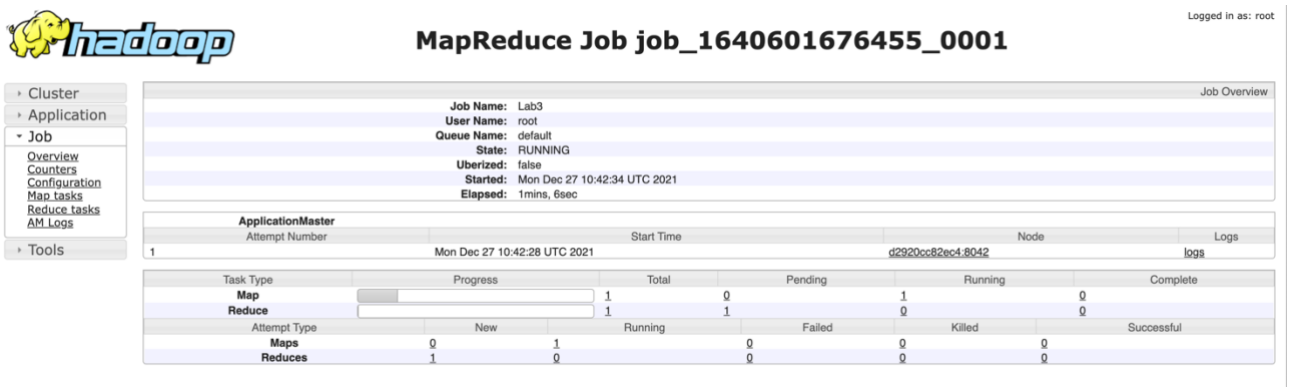
3.1Загальна концепція

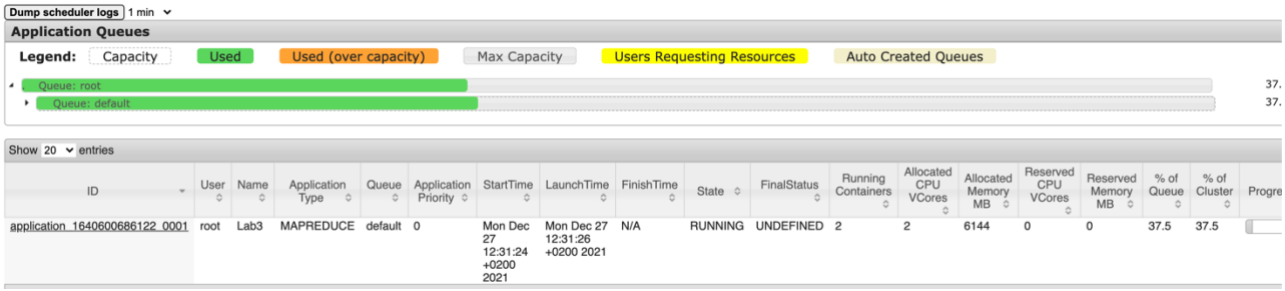
Програмне забезпечення завантажує вміст публічного Github-репозиторію у оперативну пам’ять за допомогою JGit. Далі, кожен файл даного репозиторію аналізується на предмет наявності в ньому певних патернів, що можуть свідчити про вразливість. Сам паттерн описується за допомогою регулярного виразу.

3.2Деталі роботи програмного забезпечення

Загальний принцип роботи:

1. На вхід в програму приходить шлях до директорії в HDFS, де очікується список репозиторіїв що потрібно аналізувати
2. Клас Mapper зчитує список репозиторіїв, завантажує кожен з них та записує знайдені вразливості в кожному файлі у вигляді JSON
3. Клас Reducer зчитує та поєднує всі вразливості кожного репозиторію у один JSON-список
4. На виході отримуємо пари: посилання на репозиторій, JSON зі списокм вразливостей





Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progre
application_1640600686122_0001	root	Lab3	MAPREDUCE	default	0	Mon Dec 27 12:31:24 +0200 2021	Mon Dec 27 12:31:26 +0200 2021	N/A	RUNNING	UNDEFINED	2	2	6144	0	0	37.5	37.5	

4 ОТРИМАНІ РЕЗУЛЬТАТИ

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

Всі тести проводились на локальному кластері Hadoop (процесор Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz). З стандартними налаштуваннями кластера, один nodemanager YARN одночасно може запустити 3 контейнера для задачі. Для кожного з тестів було збільшено кількість вхідних файлів з посиланнями на репозиторії.

Map tasks	**Nodes**	**Time**
1	1	2m 11s
2	1	2m 27s
4	1	4m 48s
8	1	7m 13s
4	2	3m 48s
8	2	4m 50s

5 ВИСНОВОК

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

Було порівняно результати роботи програми з різними налаштуваннями паралелізму та зі збільшенням розміру вхідних даних, можна побачити що система зможе розгорнутися навіть на великій кількості даних за умови наявності машин.

5 ПОСИЛАННЯ

[1] https://www.eclipse.org/jgit/
[2] https://github.com/big-data-europe/docker-hadoop

6 ДОДАТОК 1 - ЛОГ РОБОТИ ПРОГРАМИ

```
2021-12-27 10:31:24,715 INFO impl.YarnClientImpl: Submitted application application_1640600686122_0001
2021-12-27 10:31:24,825 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1640600686122_0001/
2021-12-27 10:31:24,828 INFO mapreduce.Job: Running job: job_1640600686122_0001
2021-12-27 10:31:40,124 INFO mapreduce.Job: Job job_1640600686122_0001 running in uber mode : false
2021-12-27 10:31:40,127 INFO mapreduce.Job:  map 0% reduce 0%
2021-12-27 10:31:59,287 INFO mapreduce.Job:  map 11% reduce 0%
2021-12-27 10:32:46,548 INFO mapreduce.Job:  map 17% reduce 0%
2021-12-27 10:32:52,587 INFO mapreduce.Job:  map 23% reduce 0%
2021-12-27 10:33:04,616 INFO mapreduce.Job:  map 28% reduce 0%
2021-12-27 10:33:10,659 INFO mapreduce.Job:  map 43% reduce 0%
2021-12-27 10:33:22,731 INFO mapreduce.Job:  map 60% reduce 0%
2021-12-27 10:33:34,782 INFO mapreduce.Job:  map 100% reduce 0%
```


```
ssue":{"issueType":"SQL dump file","issueDescription":null},"path":"2019-2020/@semester-2/week-2-2/sql-intro.sql","lineNumber":null},{"issue":{"issueType":"access_key_expose","issueDescription":"Potenti
al amazon access credential expose"},"path":"2019-2020/@semester-2/week-12-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":41},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential a
mazon access credential expose"},"path":"2019-2020/@semester-2/week-12-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":94},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazo
n access credential expose"},"path":"2019-2020/@semester-2/week-12-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":105},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon a
ccess credential expose"},"path":"2019-2020/@semester-2/week-12-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":119},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon acce
ss credential expose"},"path":"2019-2020/@semester-2/week-12-1/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":41},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access c
redential expose"},"path":"2019-2020/@semester-2/week-12-1/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":94},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access crede
ntial expose"},"path":"2019-2020/@semester-2/week-12-1/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":105},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credenti
al expose"},"path":"2019-2020/@semester-2/week-12-1/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":119},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential
expose"},"path":"2019-2020/@semester-2/week-11-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":41},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expo
se"},"path":"2019-2020/@semester-2/week-11-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":94},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"}
,"path":"2019-2020/@semester-2/week-11-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":105},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"p
ath":"2019-2020/@semester-2/week-11-2/JavaWebShop2/.idea/uiDesigner.xml","lineNumber":119}]
https://github.com/mikemelon/JavaWebEducation.git      []
https://github.com/mikemelon/java-signin.git    [{"issue":{"issueType":"Contains word: backup","issueDescription":null},"path":"backup.bat","lineNumber":null},{"issue":{"issueType":"Contains word: backu
p","issueDescription":null},"path":"backup_mysql5.5_win7.bat","lineNumber":null},{"issue":{"issueType":"SQL dump file","issueDescription":null},"path":"javaee.sql","lineNumber":null},{"issue":{"issueTyp
e":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/actions/AssignmentAction.java","lineNumber":23},{"issue":{"issueType":"acc
ess_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/actions/ManageAction.java","lineNumber":28},{"issue":{"issueType":"access_key_exp
ose","issueDescription":"Potential amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/actions/SeatSelectAction.java","lineNumber":22},{"issue":{"issueType":"access_key_expose","i
ssueDescription":"Potential amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/actions/SignInAction.java","lineNumber":31},{"issue":{"issueType":"access_key_expose","issueDescrip
tion":"Potential amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/actions/StatsAction.java","lineNumber":27},{"issue":{"issueType":"access_key_expose","issueDescription":"Poten
tial amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/actions/TaskSelectAction.java","lineNumber":21},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential am
azon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/model/AbsentRequest.java","lineNumber":14},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access
credential expose"},"path":"src/main/java/cn/lynu/lyq/signin/model/Assignment.java","lineNumber":18},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expo
se"},"path":"src/main/java/cn/lynu/lyq/signin/model/SeatAvailable.java","lineNumber":13},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":
"src/main/java/cn/lynu/lyq/signin/model/SignRecord.java","lineNumber":15},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"src/main/java/
cn/lynu/lyq/signin/model/Student.java","lineNumber":10},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"src/main/java/cn/lynu/lyq/signin
/model/Task.java","lineNumber":15},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"src/main/webapp/font/FontAwesome.otf","lineNumber":45
}]
https://github.com/sonngotung/JWebMVC.git      [{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"JWebMVCDemo/dist/JWebMVCDemo.war","line
Number":18},{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"JWebMVCDemo/dist/JWebMVCDemo.war","lineNumber":54},{"issue":{"issueType":"ac
cess_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"JWebMVCDemo/dist/JWebMVCDemo.war","lineNumber":497},{"issue":{"issueType":"access_key_expose","issueDescription":
"Potential amazon access credential expose"},"path":"JWebMVCDemo/dist/JWebMVCDemo.war","lineNumber":500},{"issue":{"issueType":"SQL dump file","issueDescription":null},"path":"JWebMVCDemo/script.sql","l
ineNumber":null}]
https://github.com/tsultana2/EducationalWebSite.git     [{"issue":{"issueType":"access_key_expose","issueDescription":"Potential amazon access credential expose"},"path":"Reading.html","lineNumber":75}]
docker run --network docker-hadoop_default --env-file hadoop.env bde2020/hadoop-base:2.0.0-hadoop3.2.1-java8 hdfs dfs -rm -r /output
```