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 3](#_Toc91059620)

[2 ВИКОРИСТАНІ БІБЛІОТЕКИ, ФРЕЙМВОРКИ 4](#_Toc91059621)

[3 ОПИС РОБОТИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ 5](#_Toc91059622)

[3.1 Загальна концепція 5](#_Toc91059623)

[3.2 Деталі роботи програмного забезпечення 5](#_Toc91059624)

[4 ОТРИМАНІ РЕЗУЛЬТАТИ 7](#_Toc91059625)

[o Вимірювання часу роботи в залежності від параметрів 7](#_Toc91059626)

[5 ВИСНОВОК 9](#_Toc91059627)

[5 ПОСИЛАННЯ 10](#_Toc91059628)

[6 ДОДАТОК 1 - ЛОГ РОБОТИ ПРОГРАМИ 11](#_Toc91059629)

# 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.

# 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].

# ОПИС РОБОТИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

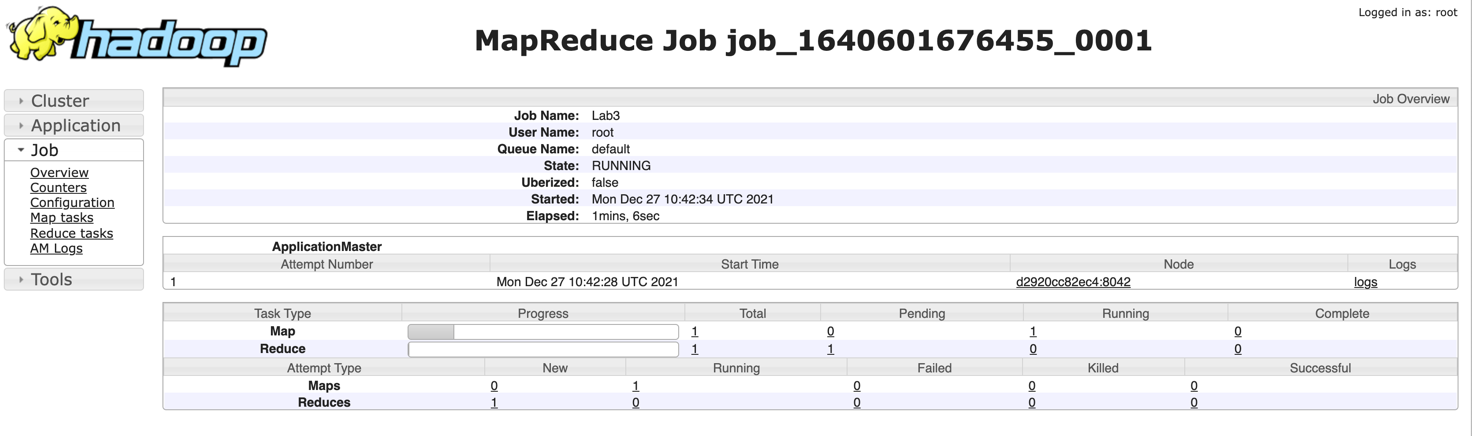
## Загальна концепція

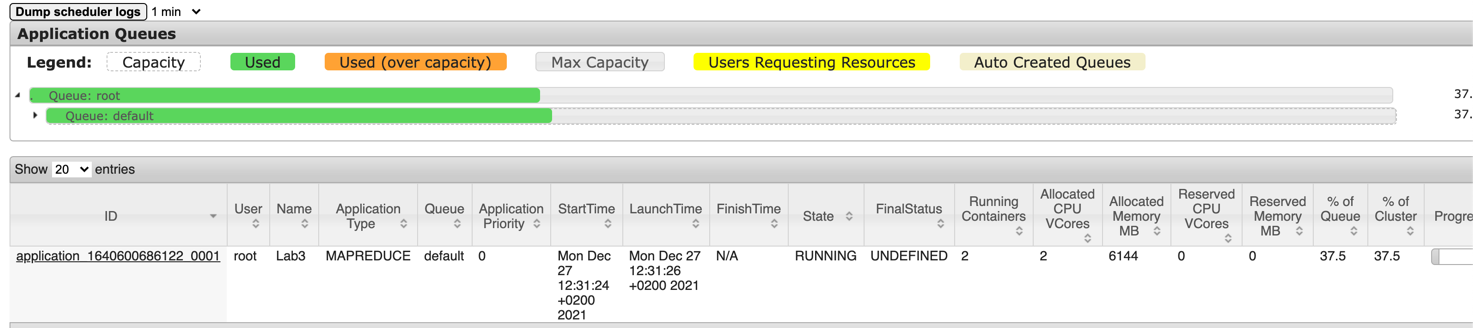
Програмне забезпечення завантажує вміст публічного Github-репозиторію у оперативну памʼять за допомогою JGit. Далі, кожен файл даного репозиторію аналізується на предмет наявності в ньому певних патернів, що можуть свідчити про вразливість. Сам паттерн описується за допомогою регулярного виразу.

## Деталі 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





# 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.

# ПОСИЛАННЯ

1. <https://www.eclipse.org/jgit/>

[2] <https://github.com/big-data-europe/docker-hadoop>

# ДОДАТОК 1 - ЛОГ РОБОТИ ПРОГРАМИ

