Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

Лабораторна робота №3

з дисципліни «Технології паралельного програмування»

за темою «Big Data з використанням засобів Apache Hadoop»

**Виконали:**

Студент групи ІС-11мн

Іванов Анатолій Ігорович

Студент групи ІА-11мн

Новиков Данило Михайлович

**Перевірив:**

вик. Жереб Константин Андрійович

Київ 2022

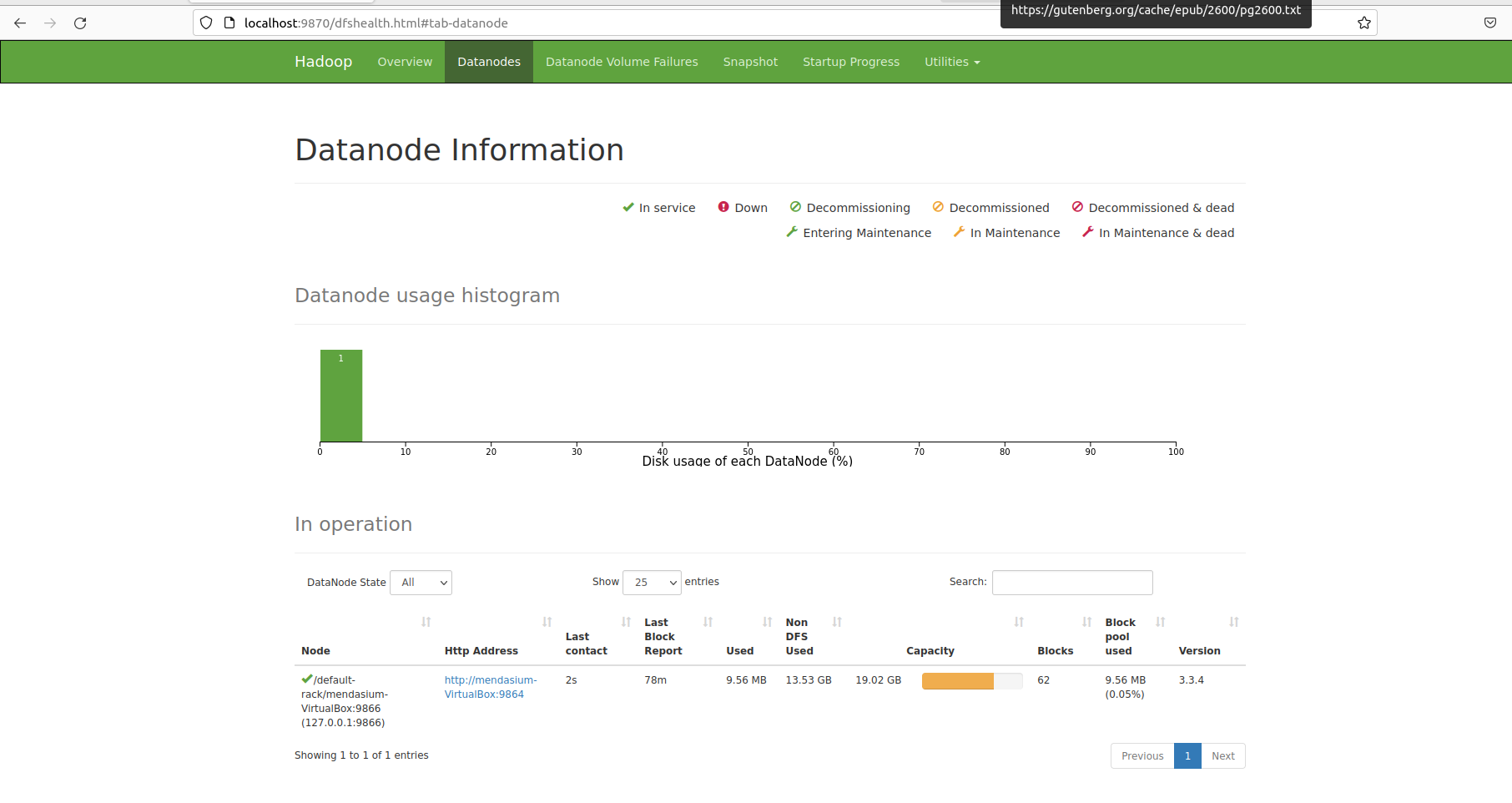
**Тема:** Big Data з використанням засобів Apache Hadoop

**Завдання:** 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.

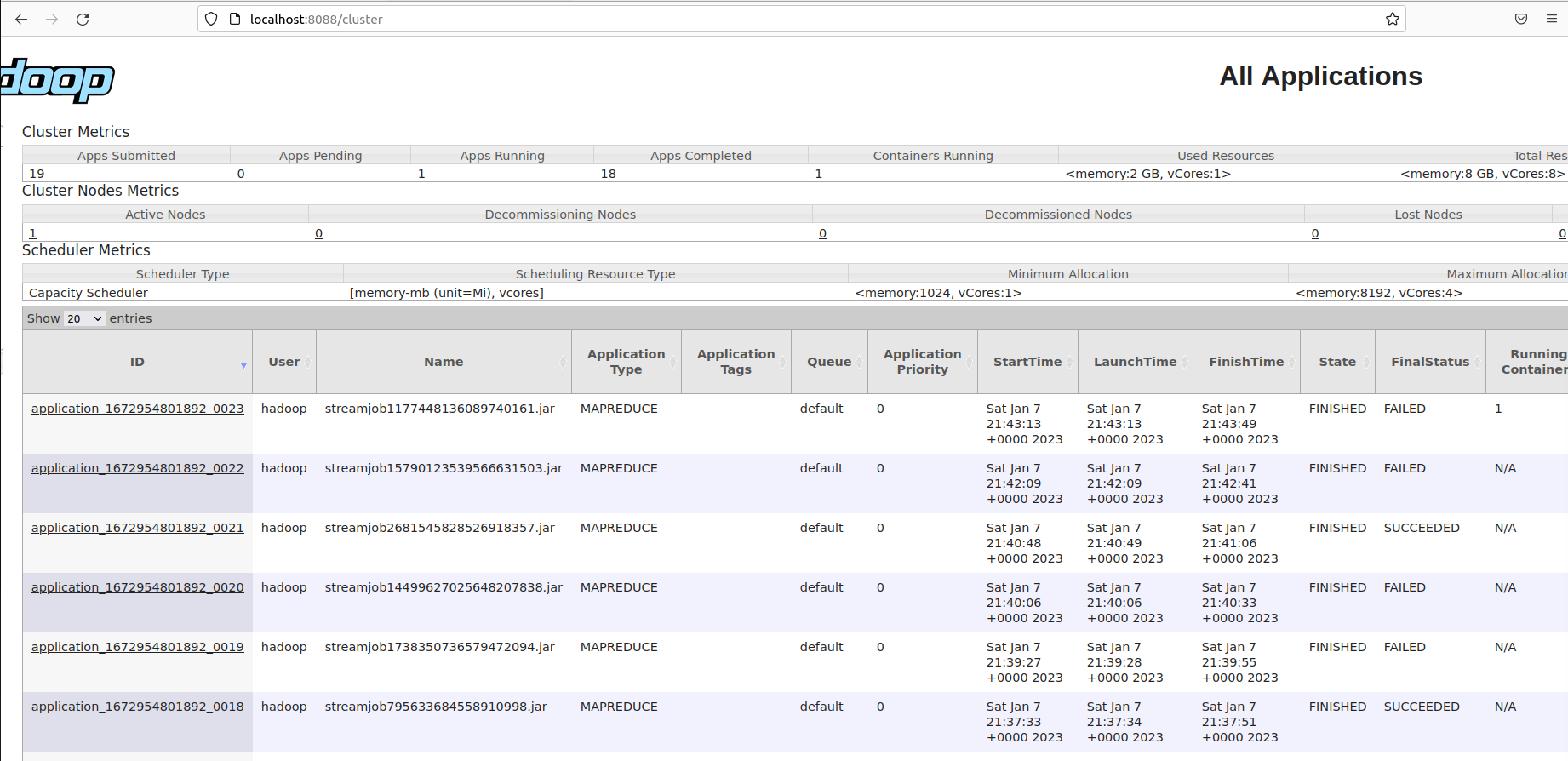
**Хід роботи**

1. Виконання даної роботи почнемо з встановдення хадупу. Для цього була створена віртуальная машина, використовуючи VirtualBox та встановлена на ній убунта. Після цього, використовуючи один з гайдів в мережі інтернет було встановлено та налаштовано Hadoop 3.3.4 – остання версія на поточний момент.
2. Після встановлення Hadoop перевіримо, що все коректно працює. Для цього перевіримо дані поточної ноди та системи в цілому.

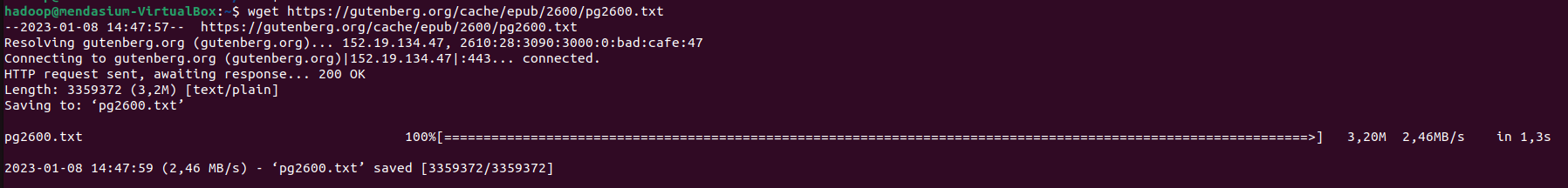
Дані поточної ноди:



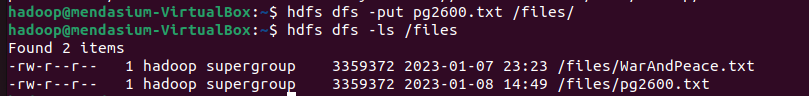
Дані роботи кластера в цілому з виконаними завданнями під час тестування системи:



1. Після встановлення та перевірки Hadoop перейдемо до створення програми map-reduce, що буде рахувати частоту входження слів у тексті/файлі та виводити результат даної частоти. Для цього використаємо Hadoop-streaming та python як мову програмування.
2. По-перше завантажимо файл, що будемо аналізувати за допомогою команди wget. В даному випадку був обраний роман Війна та Мир у тестовому файлі, в якому присутні всі томи. Посилання: <https://www.gutenberg.org/cache/epub/2600/pg2600.txt>



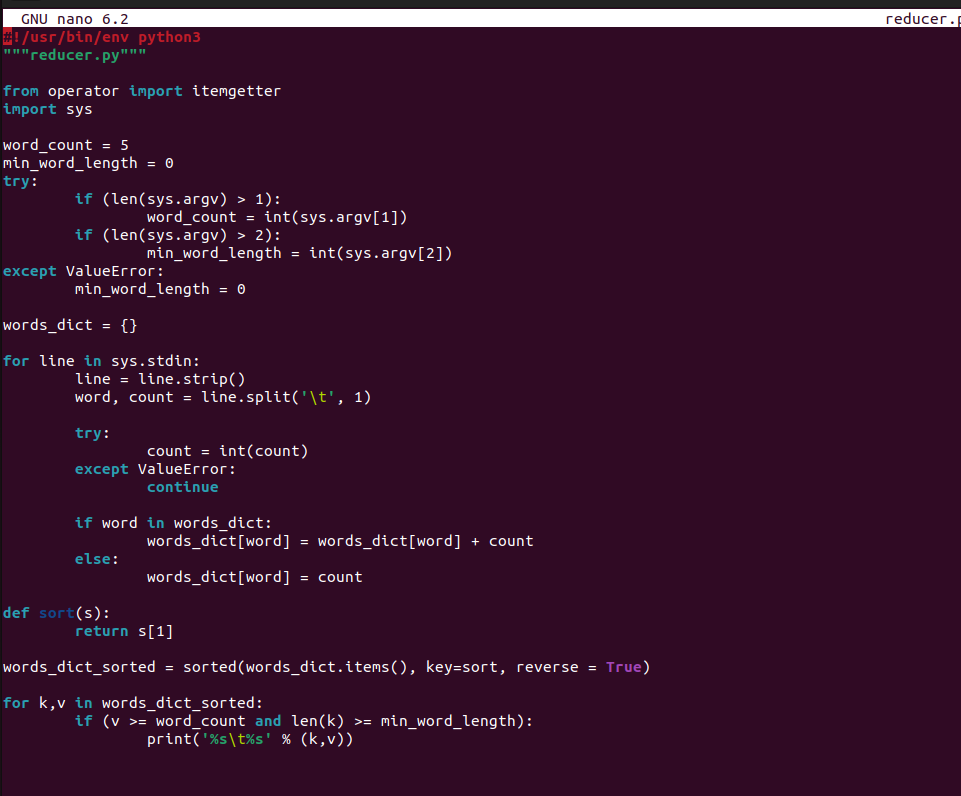
1. Після завантаження, додамо даний файл до файлової системи Hadoop та перевіримо його наявність.



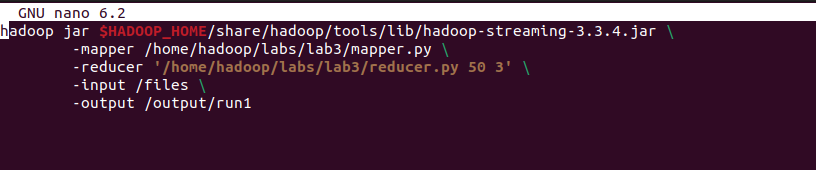
1. Перейдемо до написання коду та створимо mapper та reducer файли, які будуть рахувати кількість входжень слів в текстовому файлі. У маппері розіб’ємо текст на слова та видалимо усі зайві символи (.,?!»).



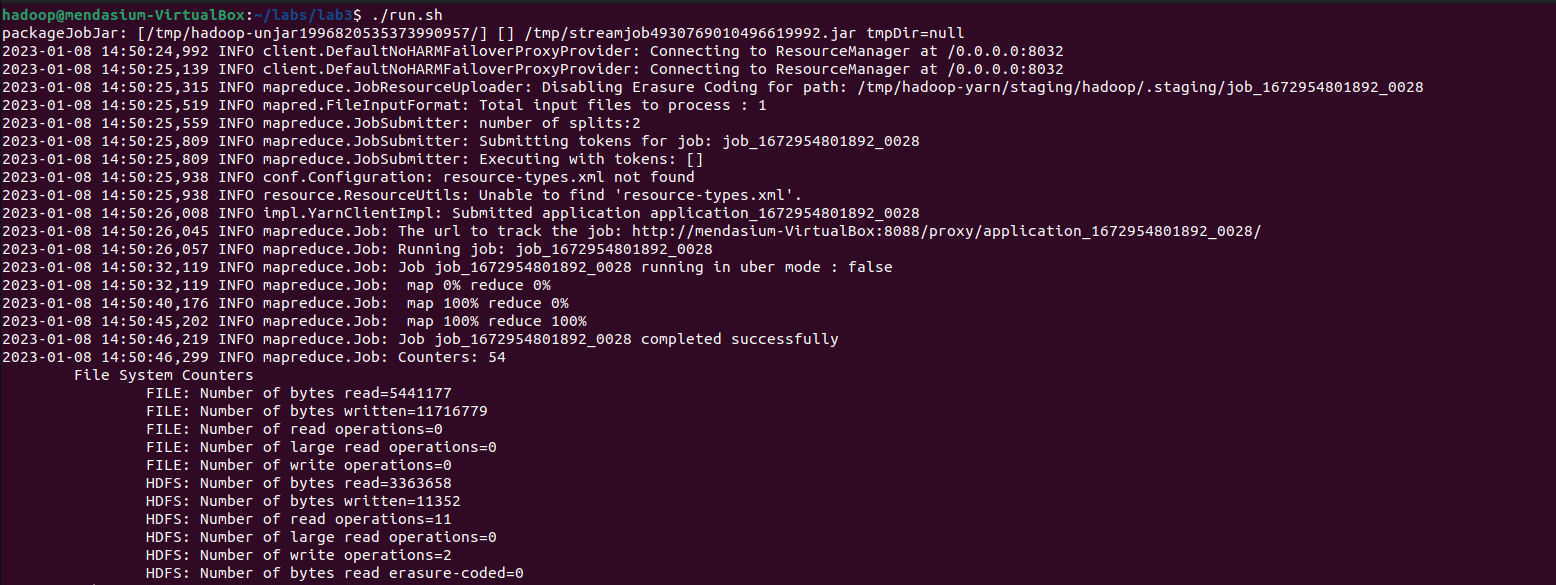
1. Маппер повертає дані у форматі (word, 1), де 1 – кількість входження слова для більш легкого підрахунку. У reducer створимо словник цих слів та порахуємо їх частоту входження. Після цього відсортуємо слова за частотою входження. При цьому, були додані параметри мінімальної довжини слова, щоб відкинути прийменники та мінімальну кількість входжень для додавання до результату.

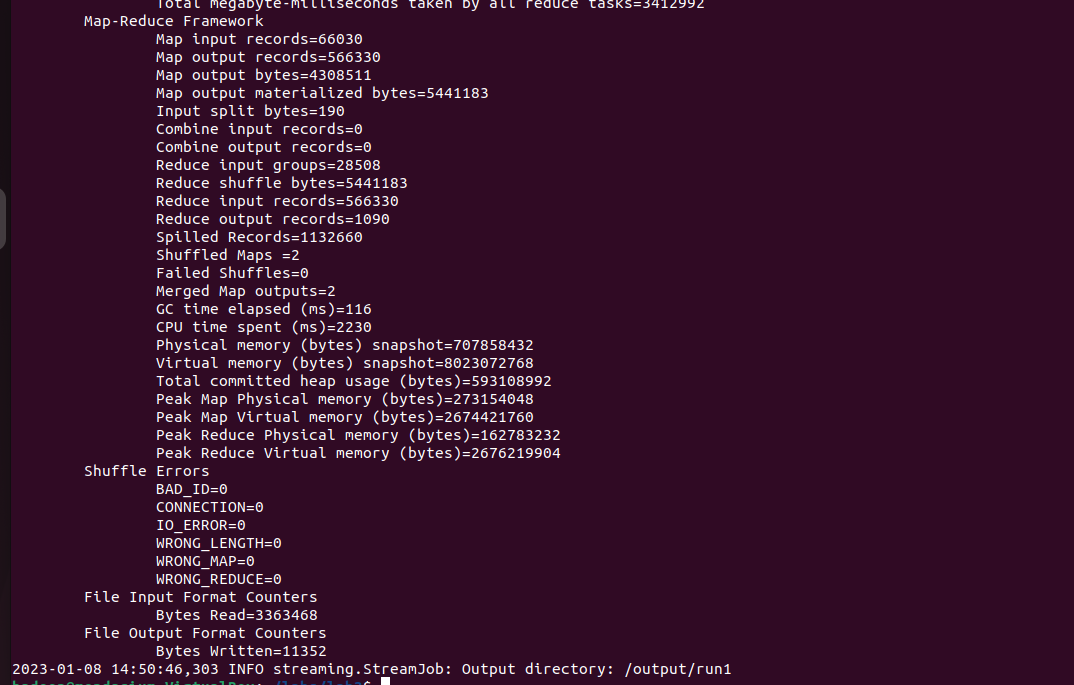


1. Після створення даних файлів, запустимо їх за допомогою hadoop-streaming. Reducer передамо параметри 50 та 3.



1. Запустимо отриману команду та отримаємо наступний результат виконання.





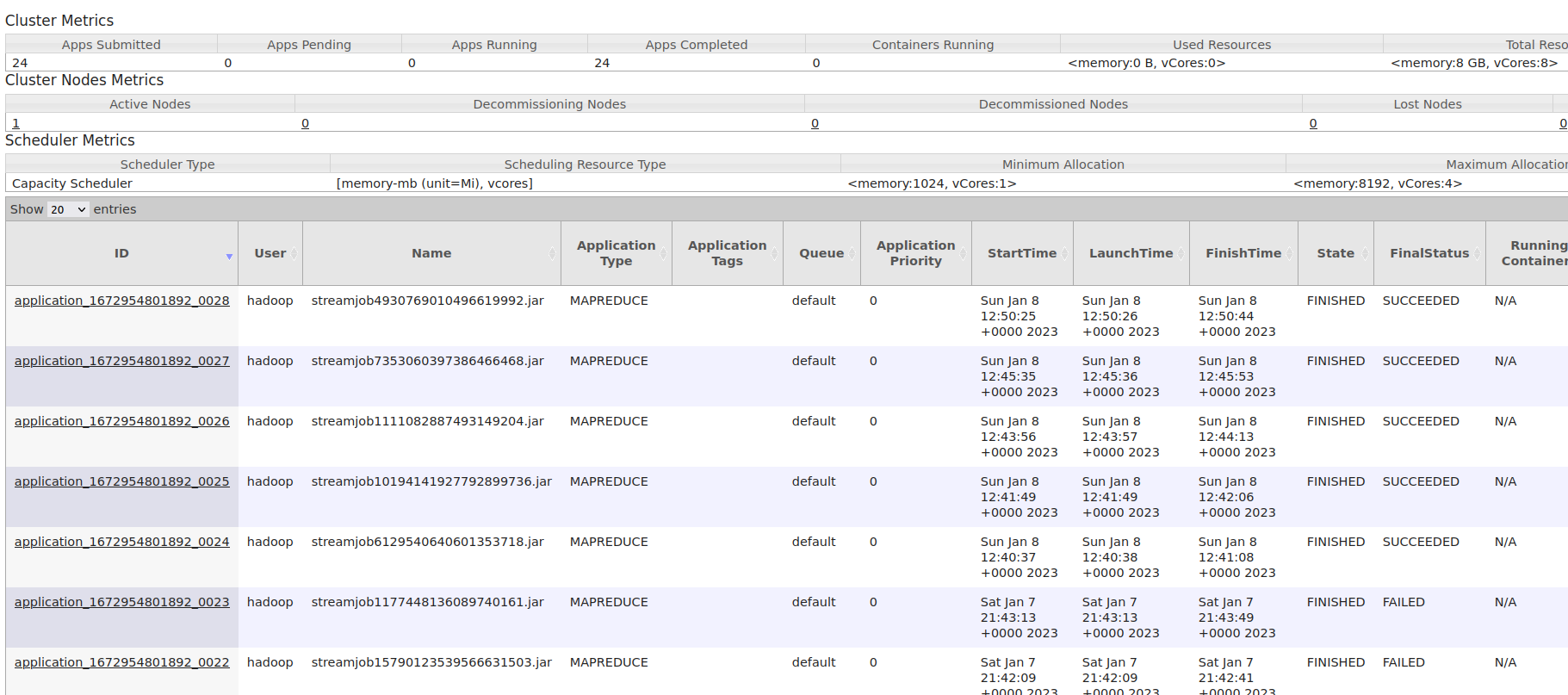
1. Оскільки Hadoop зберігає результат та не виводить його під час виконання, переглянемо його за допомогою наступної команди.



1. Отриманий результат виконання написаної програми:

|  |  |
| --- | --- |
|  |  |

1. Можемо побачити, що найбільше входження отримали слова “the”, “and”, “his” та “was”. З більш значимих слів маємо слово said з кількостю входжень 2636.
2. Також, під час запуску додаються нові запущенні та виконані застосунки у кластері.



Посилання на репозиторій: <https://github.com/JokerFunny/parallel_programming>

**Висновок:**

в результаті виконання даної лабораторної роботи було запущено Hadoop та створену просту програму для підрахунку кількості входжень слів у тексті.