

Побудова моделей обробки великих масивів даних з використанням технології MapReduce

Автор роботи:

Студент групи ДА-52м

Яременко Вадим Сергійович

Об'єкт дослідження

Великі масиви даних

Предмет дослідження

Моделі для вирішення задач кластеризації та пошуку частих предметних наборів, які основані на технології MapReduce

Мета та завдання

- Метою даної роботи є дослідження технології MapReduce та її використання для вирішення задач обробки великих масивів даних на прикладі задач кластеризації та пошуку частих предметних наборів.
- Результатом проведених досліджень є практична частина роботи, що становить собою створення моделей для обробки великих даних та їх апробація з використанням сучасних програмних засобів.

Задача кластеризації

Дано: X – простір об'єктів;
 $X_l = \{x_i\}_{i=1}^l$ – вибірка елементів;
 $d: X \times X \rightarrow [0; \infty)$ – функція відстані між об'єктами

Знайти: Y – множину кластерів і відображення $a: X \rightarrow Y$ – алгоритм кластеризації такий, що кожен кластер складається з близьких між собою об'єктів, а об'єкти різних кластерів суттєво відрізняються

Алгоритм CURE

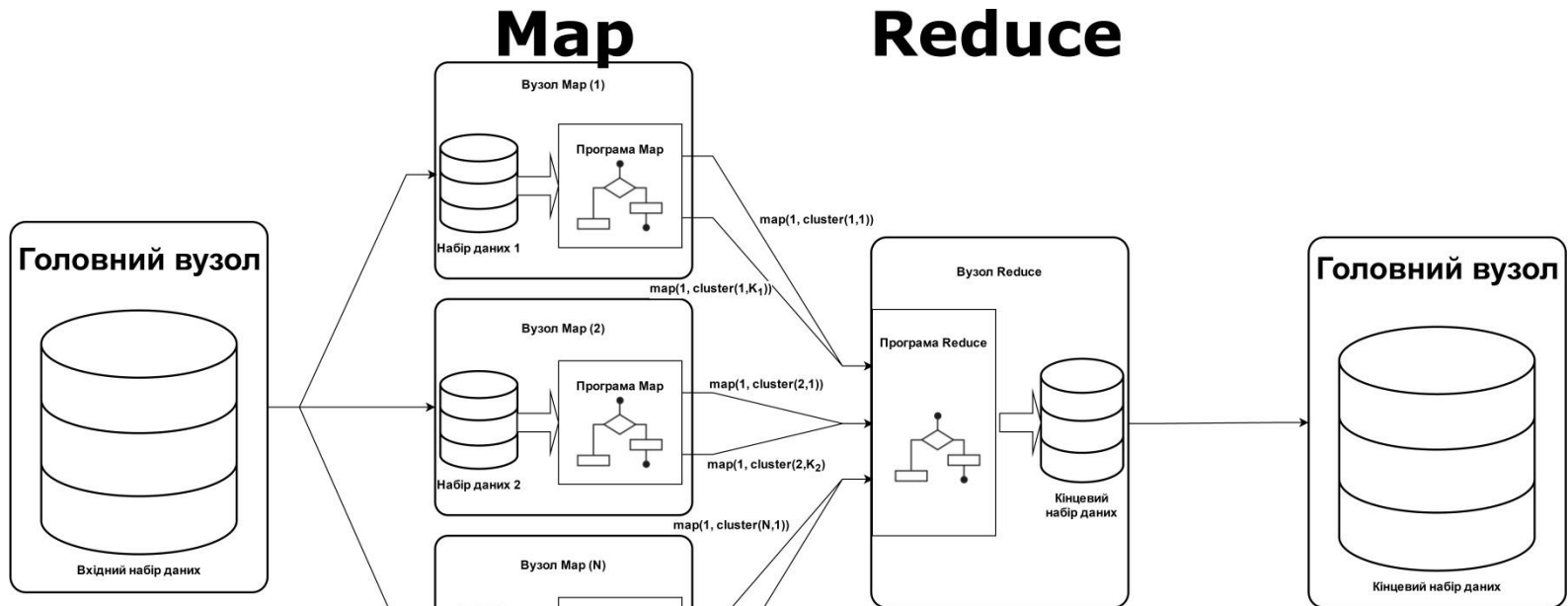
Перший етап

1. Обрати частину вибірки і у оперативній пам'яті провести кластеризацію
2. Обрати репрезентативні точки
3. Змістити репрезентативні точки у напрямку центроїду кластеру на фіксований відсоток відстані

Другий етап

1. Об'єднати кластери, репрезентативні точки яких є достатньо близькими

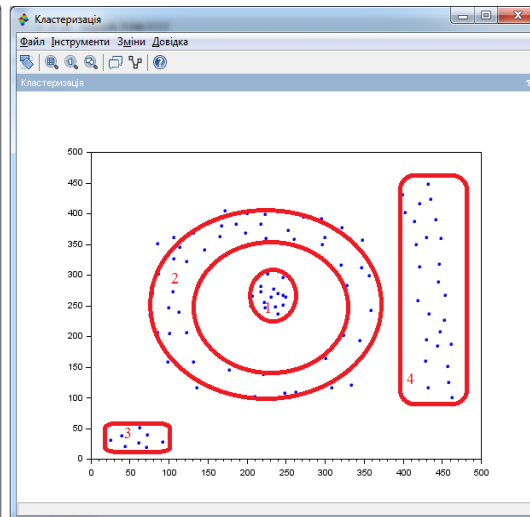
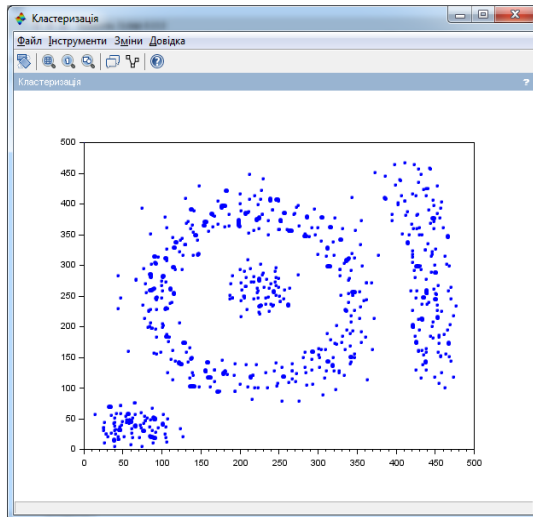
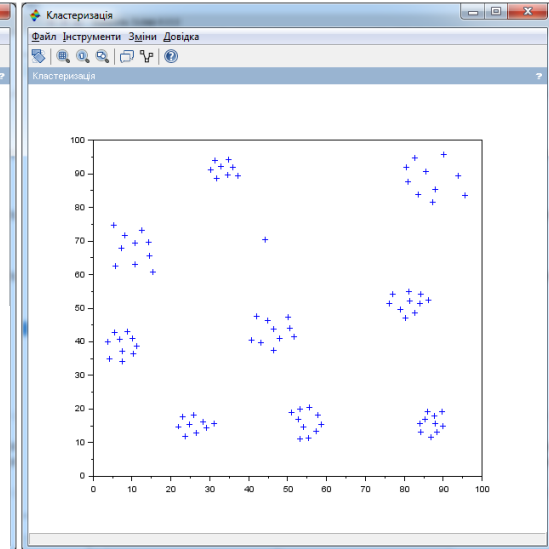
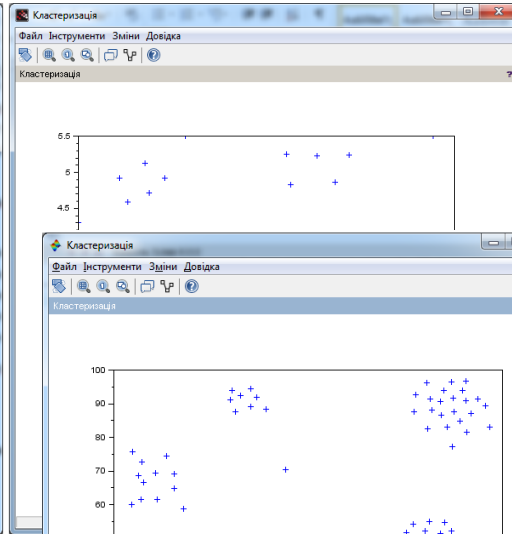
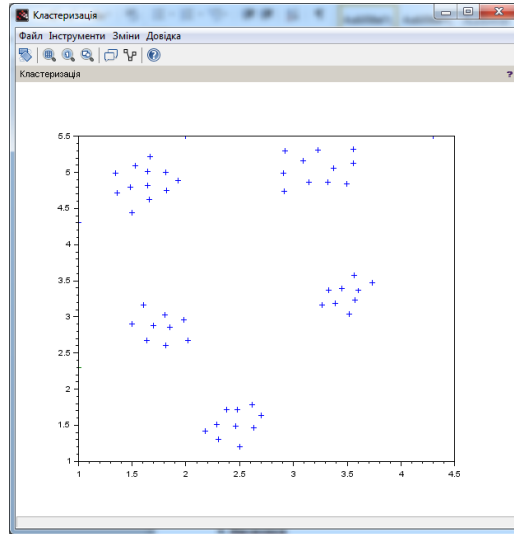
Модель для рішення задачі кластеризації з використанням MapReduce



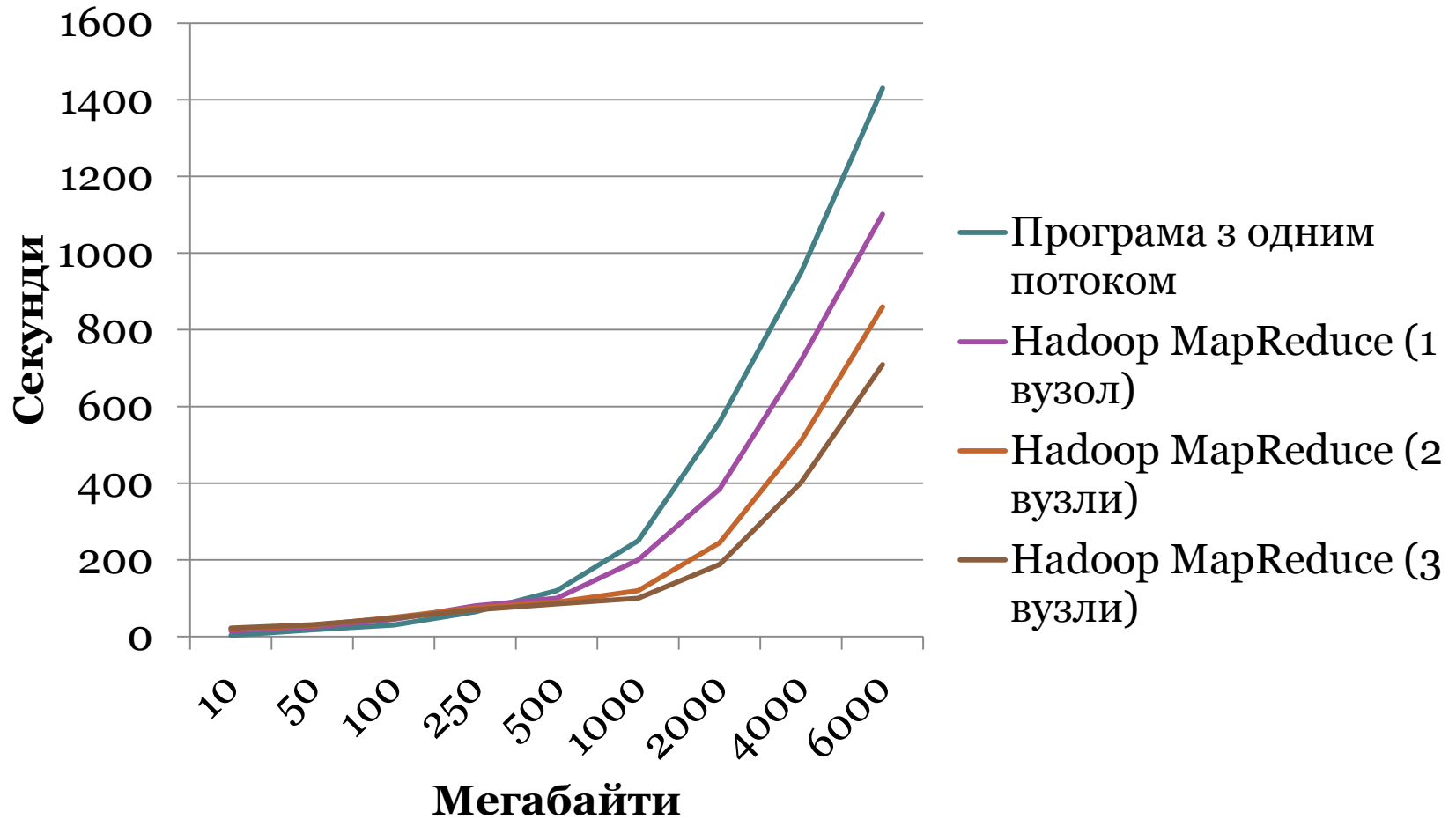
**Репрезентативні
виборки
кластерів, які
знайдені на
кожному з
обчислювальних
вузлів**

**Репрезентативні
виборки усіх
знайдених
кластерів**

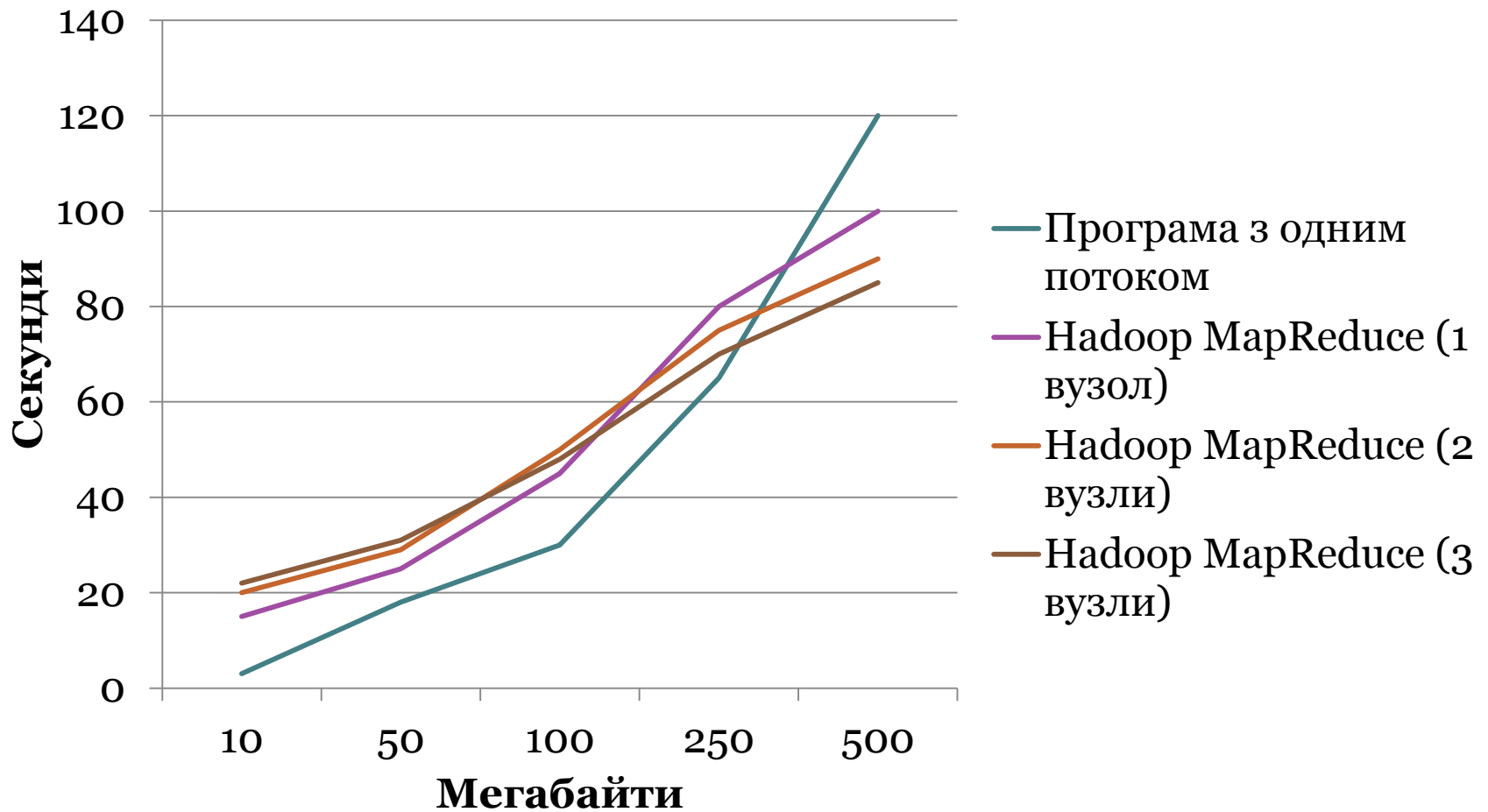
Тестування моделі



Аналіз швидкості роботи



Аналіз швидкості роботи (вхідні дані менші за 500 МБ)



Задача пошуку частих предметних наборів

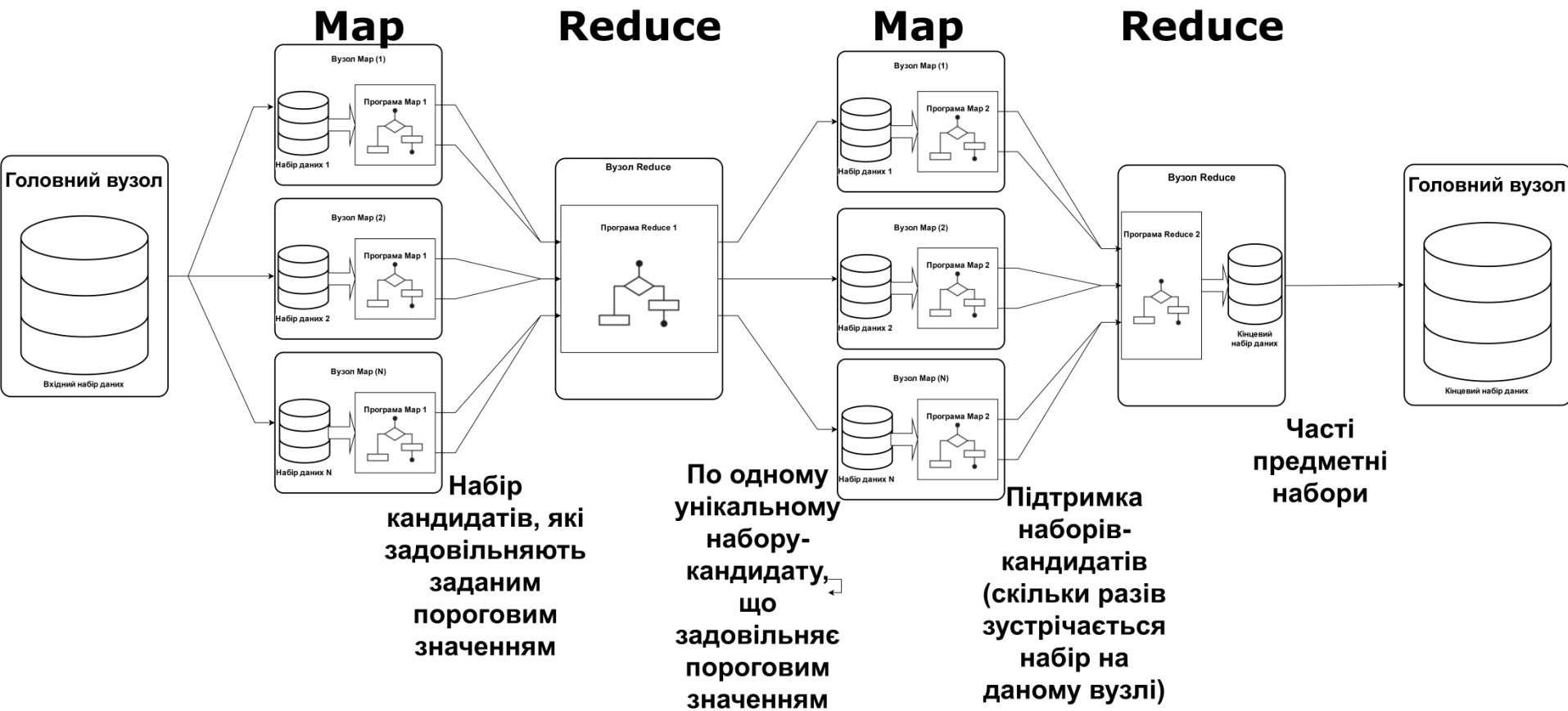
Дано: Множина кошиків з набором предметів
 S – порогова підтримка
 I – підмножина предметів
 K – підтримка I – кількість кошиків, для яких I є підмножиною

Знайти: $\{I\}$ – підмножини, для яких підтримка не менша S

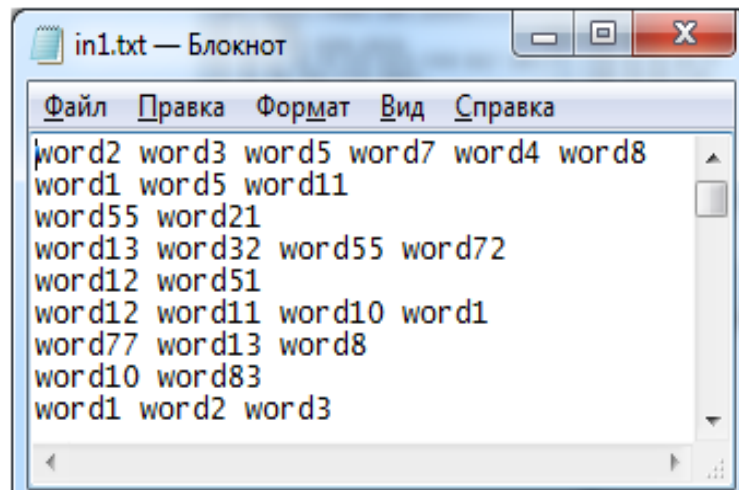
Алгоритм SON

1. Розділяємо вхідну вибірку даних на частини
2. Шукаємо часті набори в кожній з отриманих частин виборки
3. Робимо ще один прохід по даним та рахуємо скільки разів кожен із знайдених наборів зустрічається у загальній виборці та інкрементуємо кількість входжень
4. Обираємо частими наборами ті, у яких кількість входжень є найбільша

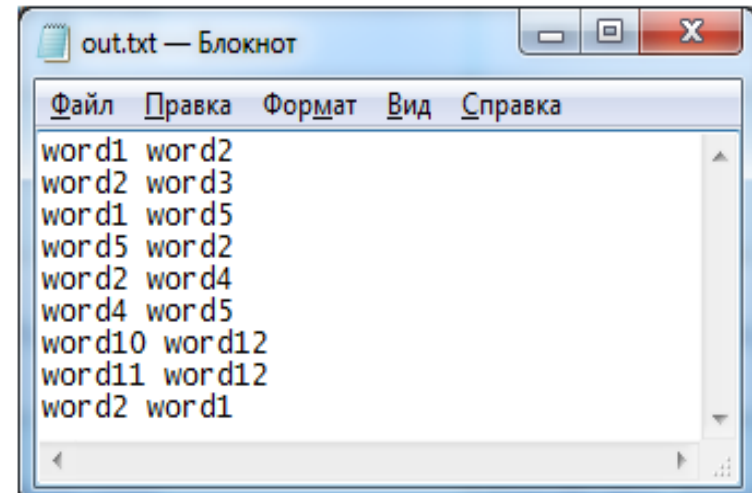
Модель для рішення задачі пошуку частих предметних наборів з використанням MapReduce



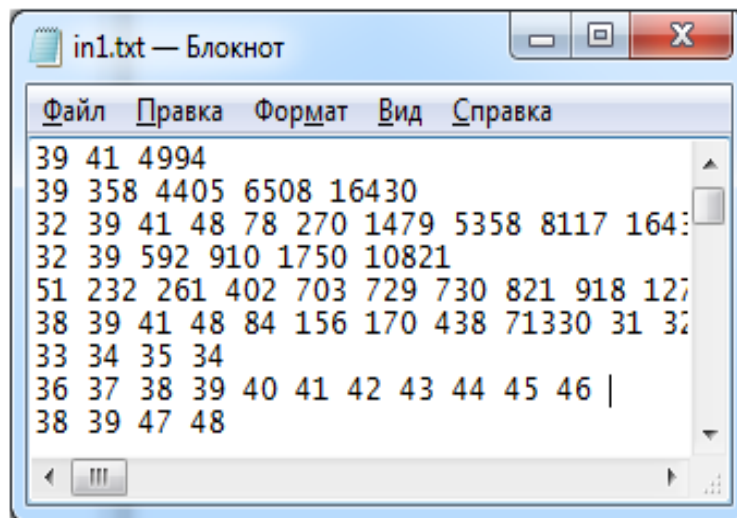
Тестування моделі



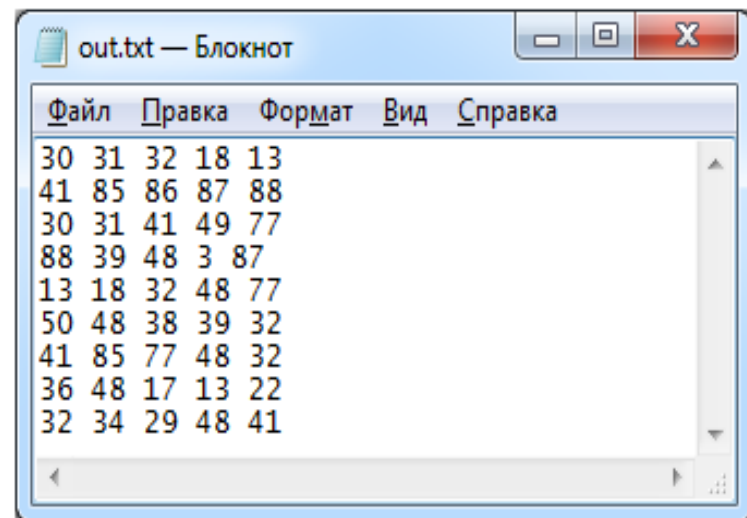
```
in1.txt — Блокнот
Файл  П_равка  Ф_ормат  В_ид  С_правка
word2 word3 word5 word7 word4 word8
word1 word5 word11
word55 word21
word13 word32 word55 word72
word12 word51
word12 word11 word10 word1
word77 word13 word8
word10 word83
word1 word2 word3
```



```
out.txt — Блокнот
Файл  П_равка  Ф_ормат  В_ид  С_правка
word1 word2
word2 word3
word1 word5
word5 word2
word2 word4
word4 word5
word10 word12
word11 word12
word2 word1
```

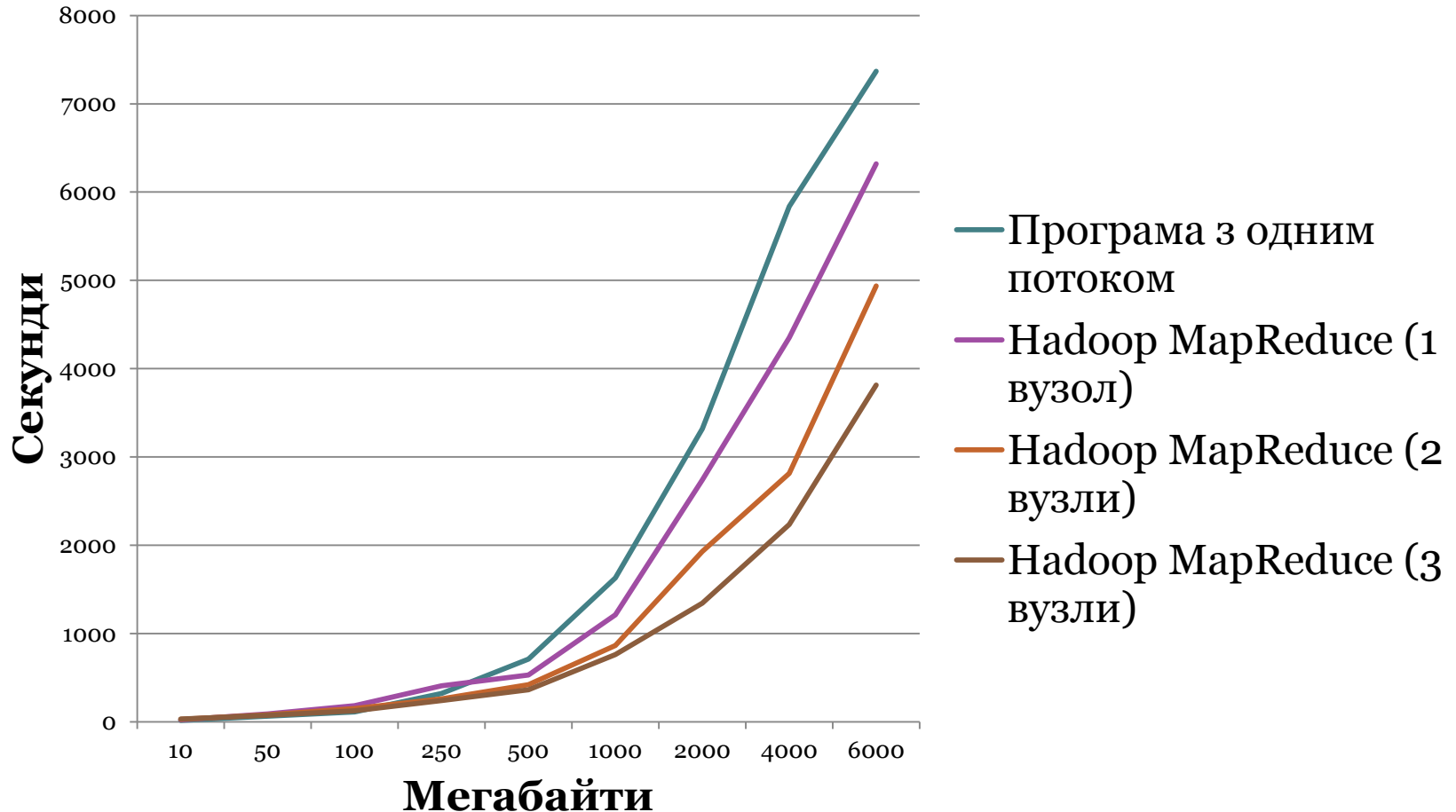


```
in1.txt — Блокнот
Файл  П_равка  Ф_ормат  В_ид  С_правка
39 41 4994
39 358 4405 6508 16430
32 39 41 48 78 270 1479 5358 8117 16430
32 39 592 910 1750 10821
51 232 261 402 703 729 730 821 918 1270
38 39 41 48 84 156 170 438 71330 31 32
33 34 35 34
36 37 38 39 40 41 42 43 44 45 46 |
38 39 47 48
```

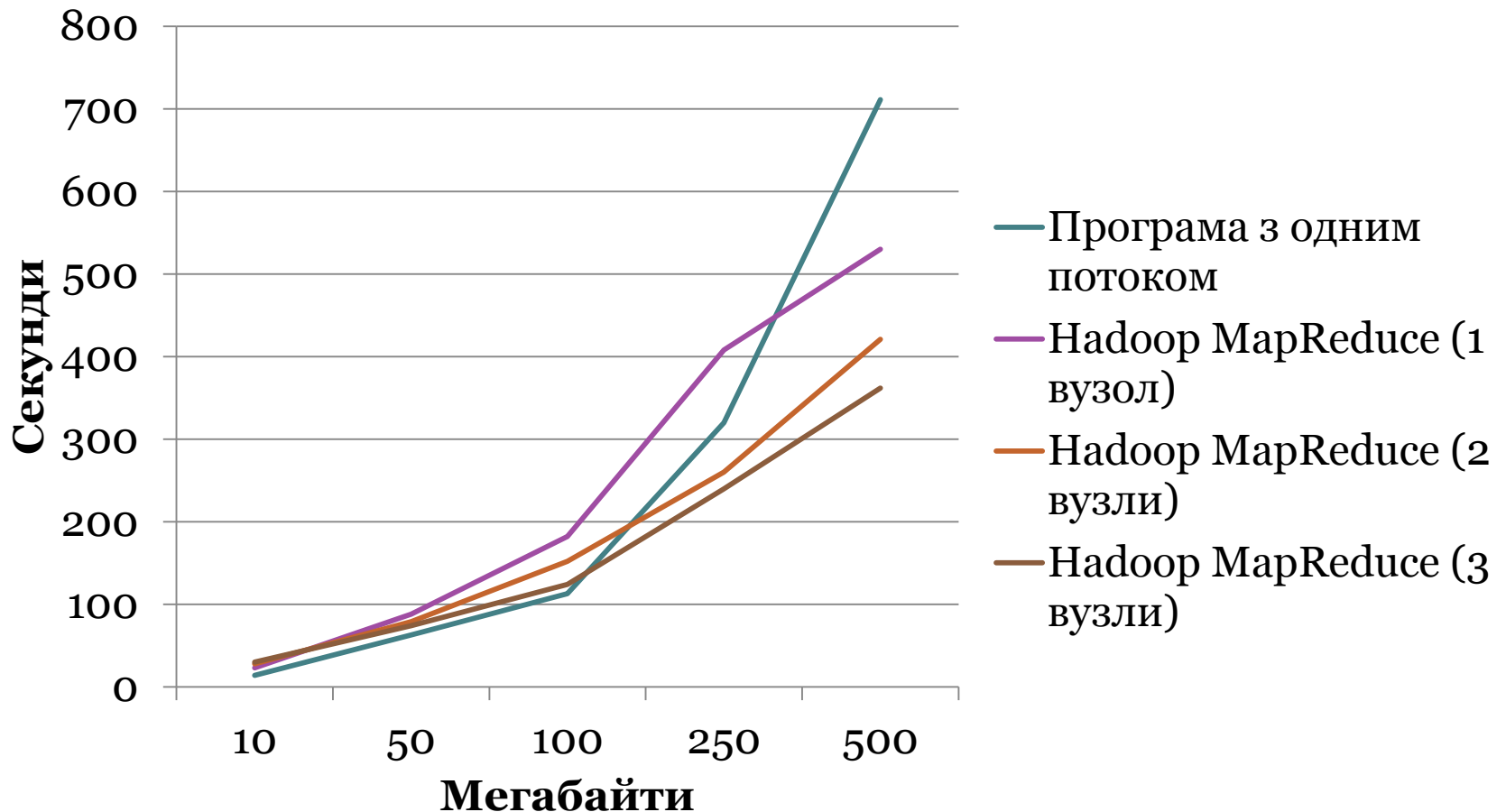


```
out.txt — Блокнот
Файл  П_равка  Ф_ормат  В_ид  С_правка
30 31 32 18 13
41 85 86 87 88
30 31 41 49 77
88 39 48 3 87
13 18 32 48 77
50 48 38 39 32
41 85 77 48 32
36 48 17 13 22
32 34 29 48 41
```

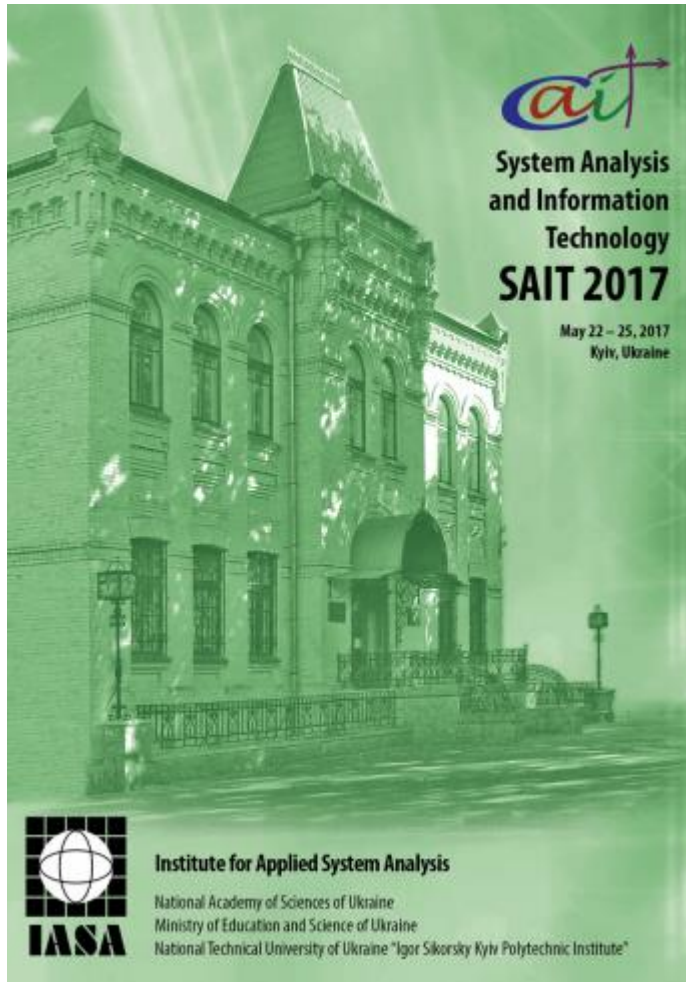
Аналіз швидкості роботи



Аналіз швидкості роботи (вхідні дані менші за 500 МБ)



Публікації



МІЖНАРОДНИЙ НАУКОВИЙ ЖУРНАЛ «ІНТЕРНАУКА»

ISSN 2520-2057



INTERNATIONAL
SCIENTIFIC JOURNAL
«INTERNAUKA»

МЕЖДУНАРОДНЫЙ
НАУЧНЫЙ ЖУРНАЛ
«ИНТЕРНАУКА»

№ 6 (28) / 2017



Майбутні напрямки роботи та досліджень

- Викласти вихідні коди  **GitHub**
- Дослідити роботу моделі для вирішення задачі кластеризації в залежності від розподілення вхідних даних між вузлами
- Дослідити технологію  **Spark**
- Дослідити поточний стан розвитку алгоритмів обробки великих даних з використанням квантових комп'ютерів



Дякую за увагу!

Студент групи ДА-52м
Яременко Вадим Сергійович