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;
$d: X \times X \to [0; \infty)$ – функція відстані між об’єктами

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

Алгоритм CURE

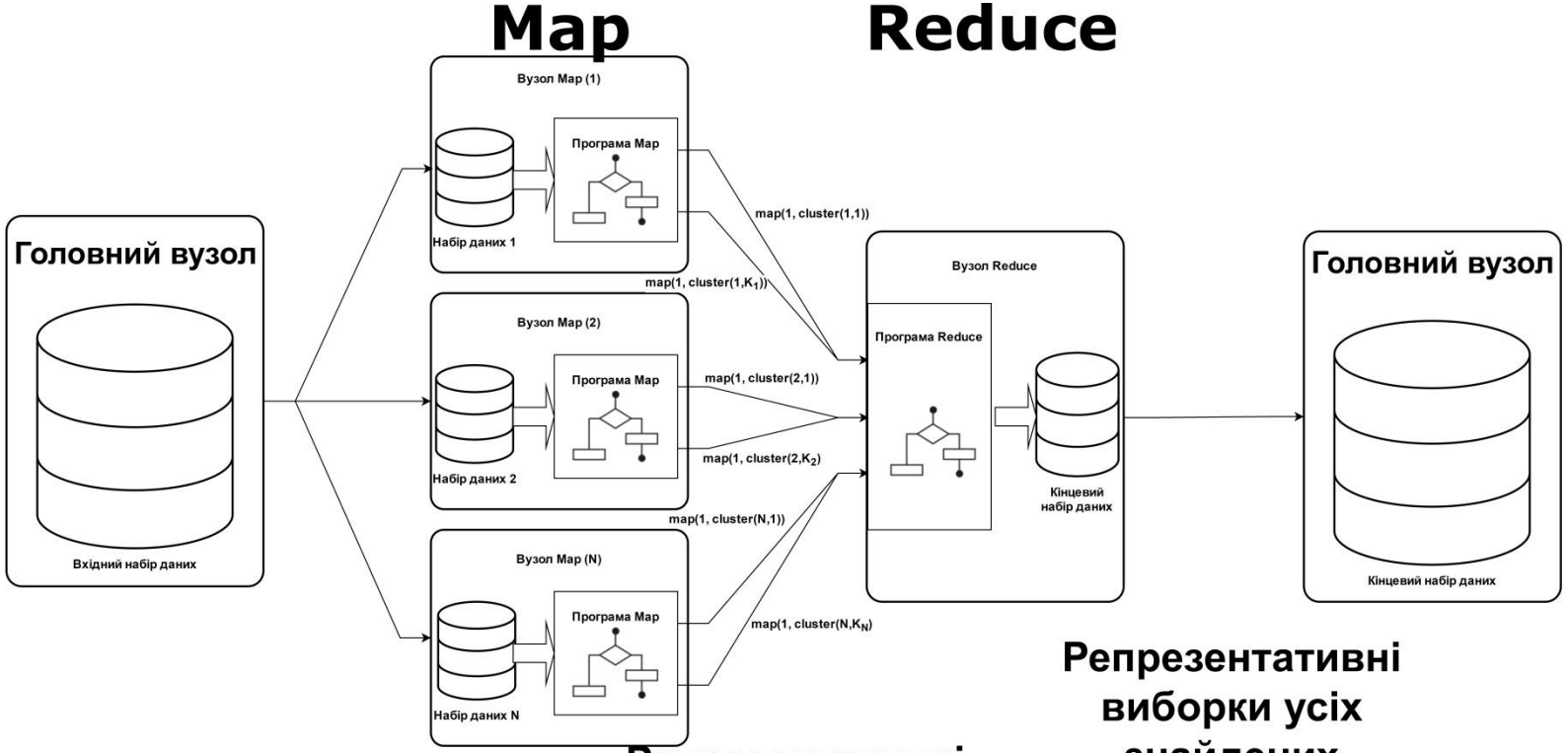
Перший етап

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

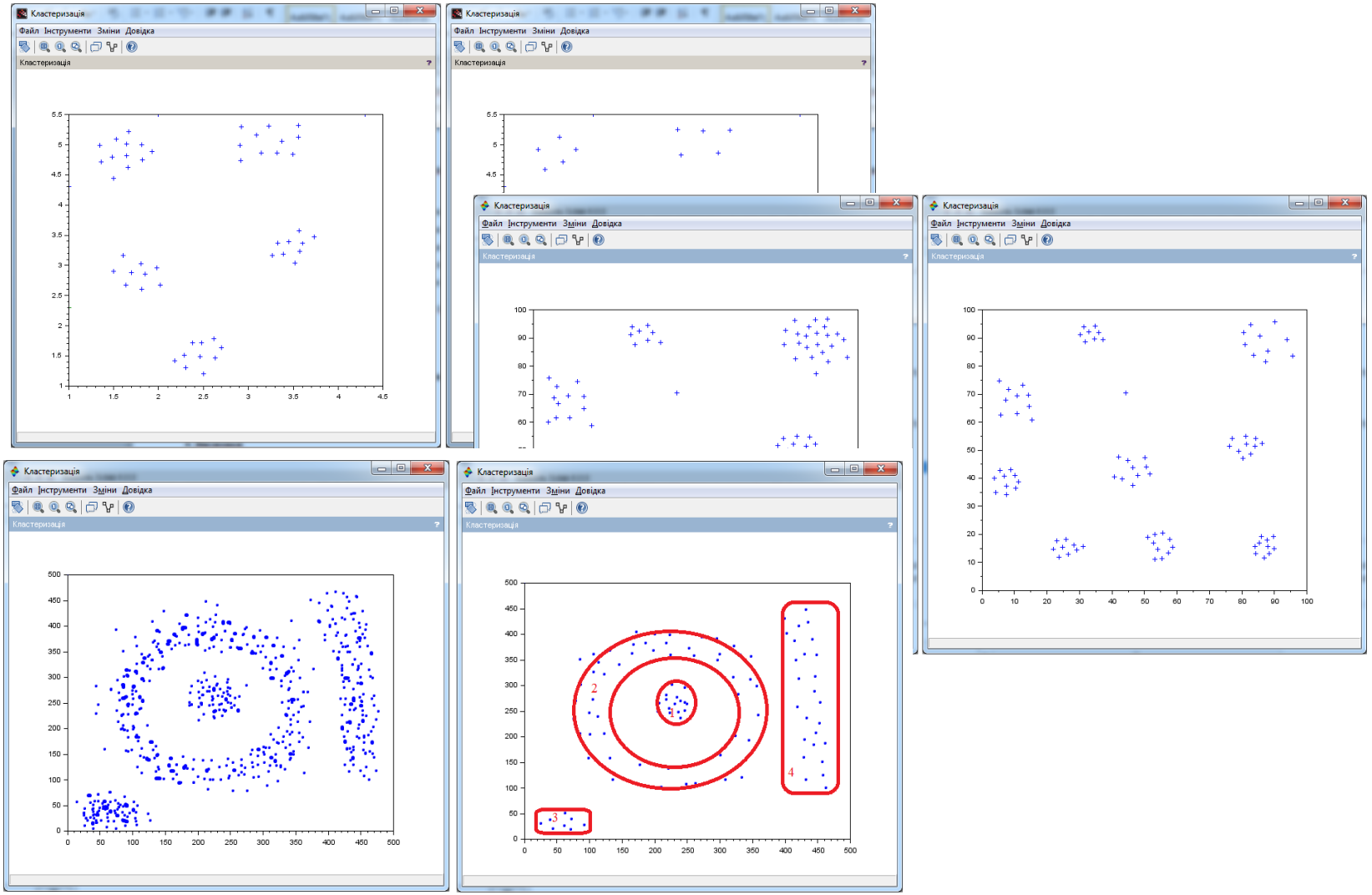
Другий етап

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

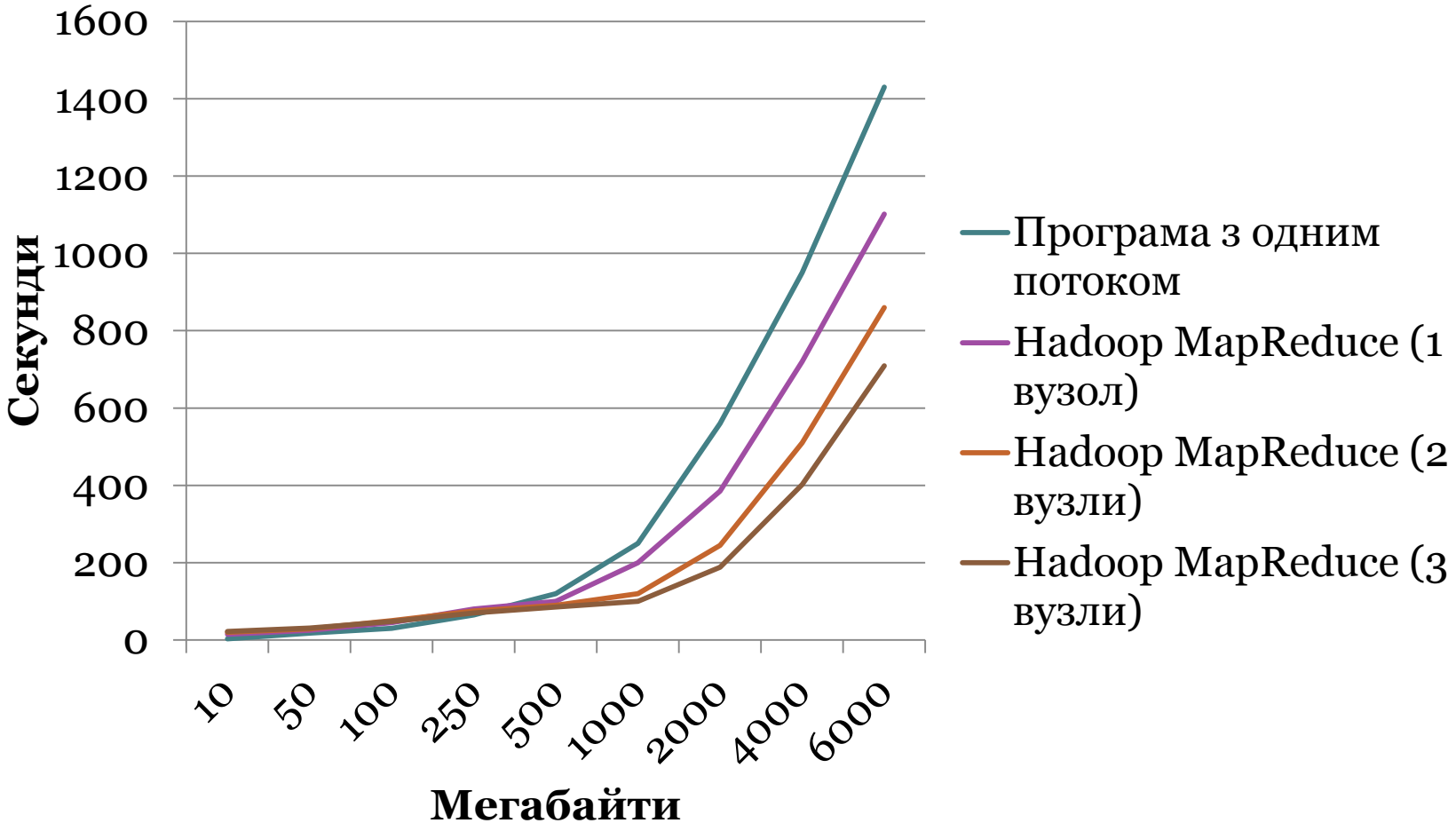
Модель для рішення задачі кластеризації з використанням MapReduce



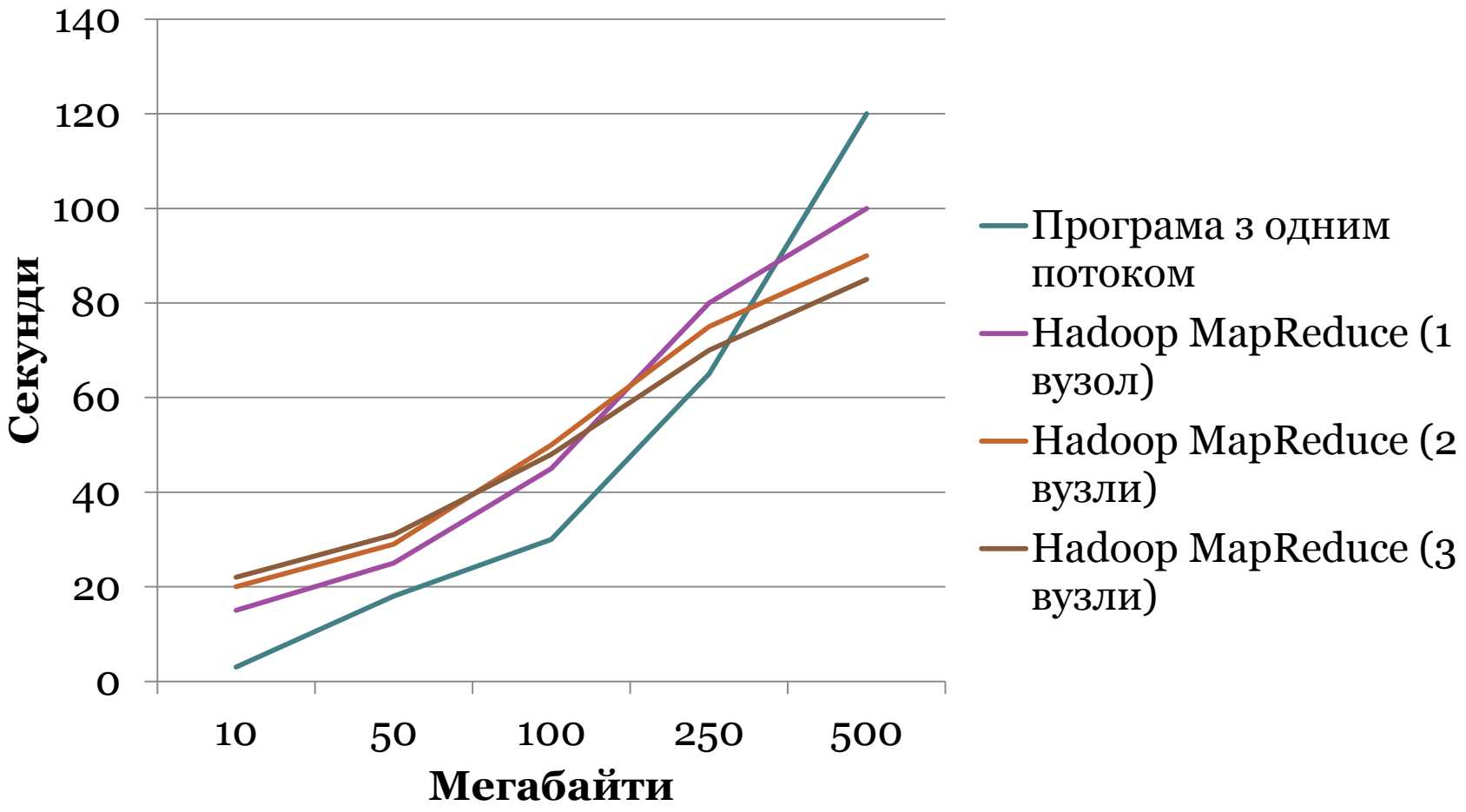
Тестування моделі



Аналіз швидкості роботи



Аналіз швидкості роботи (вхідні дані менші за 500 МБ)



Задача пошуку частих предметних наборів

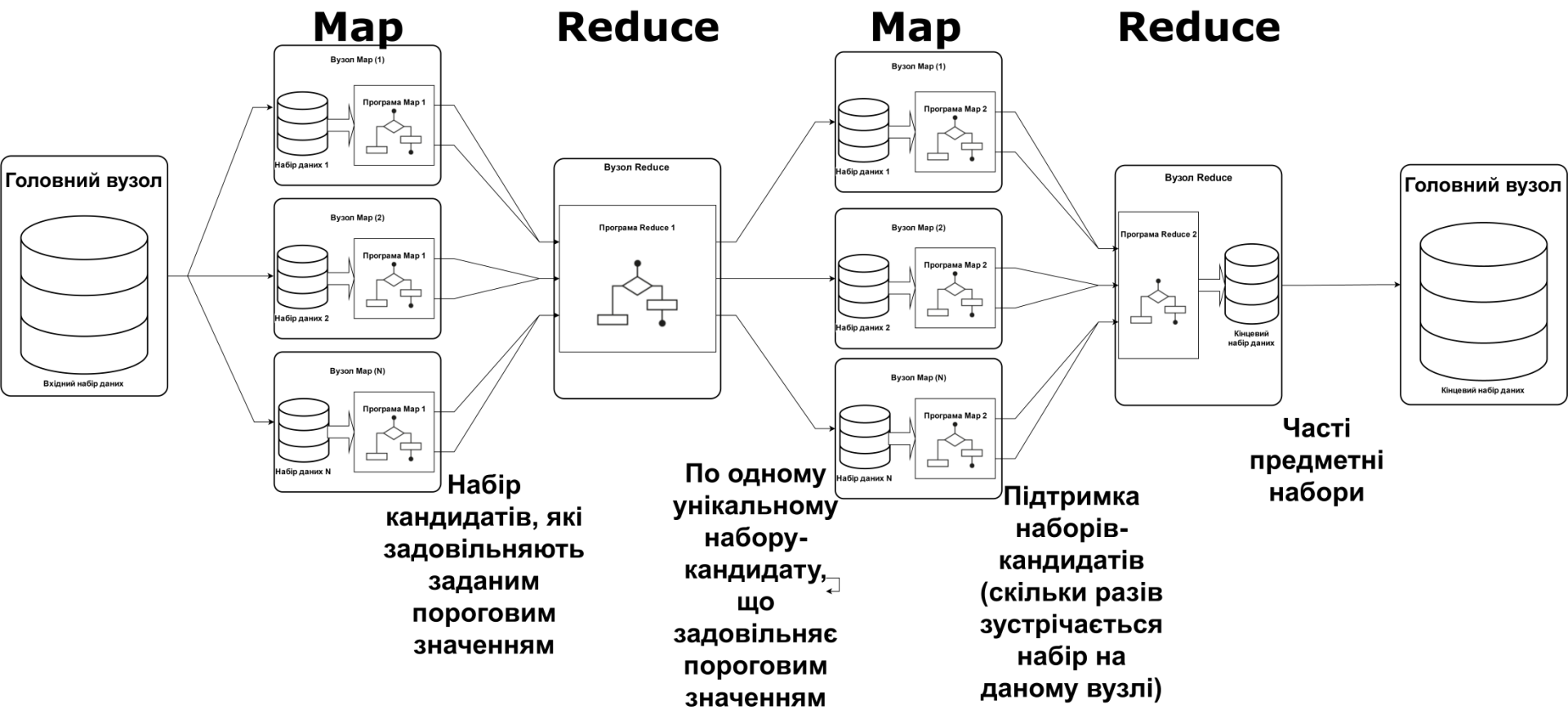
Дано: Множина кошиків з набором предметів
S – порогова підтримка
I – підмножина предметів
K – підтримка I – кількість кошиків, для яких I є підмножиною

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

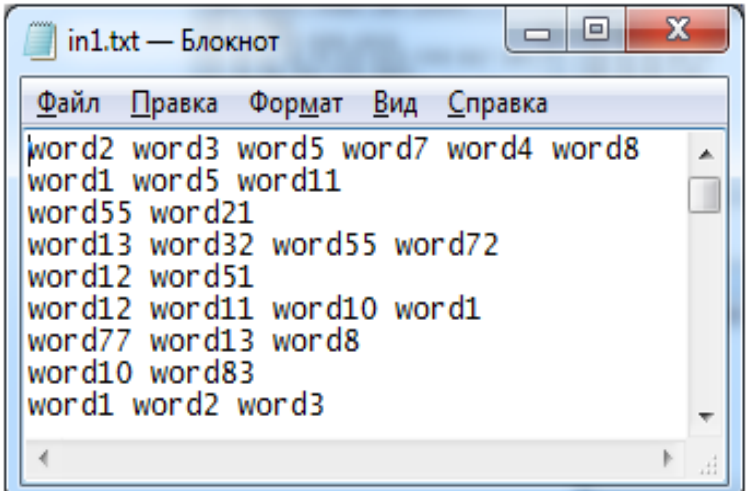
Алгоритм SON

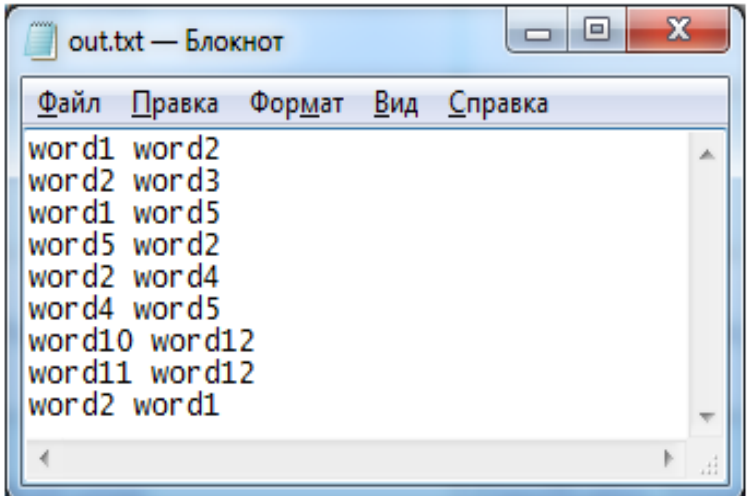
1. Розділяємо вхідну вибірку даних на частини
2. Шукаємо часті набори в кожній з отриманих частин виборки
3. Робимо ще один прохід по даним та рахуємо скільки разів кожен із знайдених наборів зустрічається у загальній виборці та інкрементуємо кількість входжень
4. Обираємо частими наборами ті, у яких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

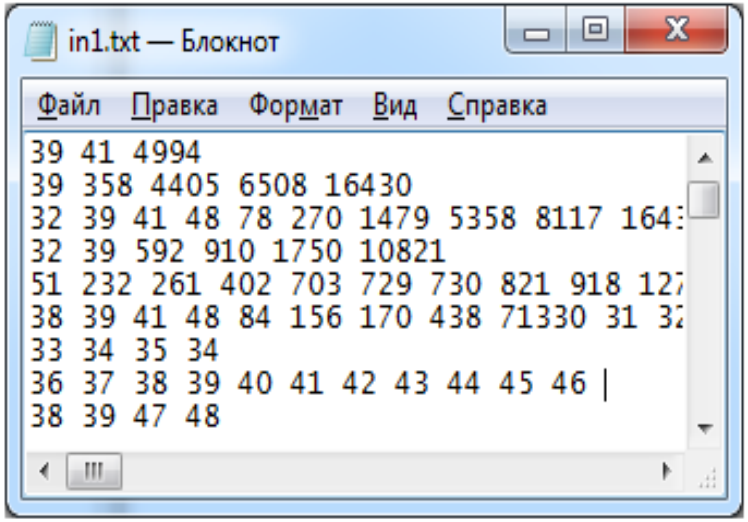
Модель для рішення задачі пошуку частих предметних наборів з використанням MapReduce

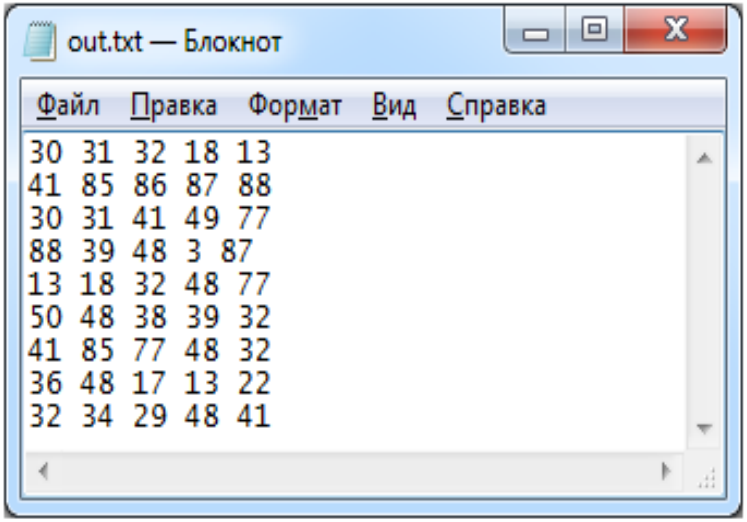


Тестування моделі

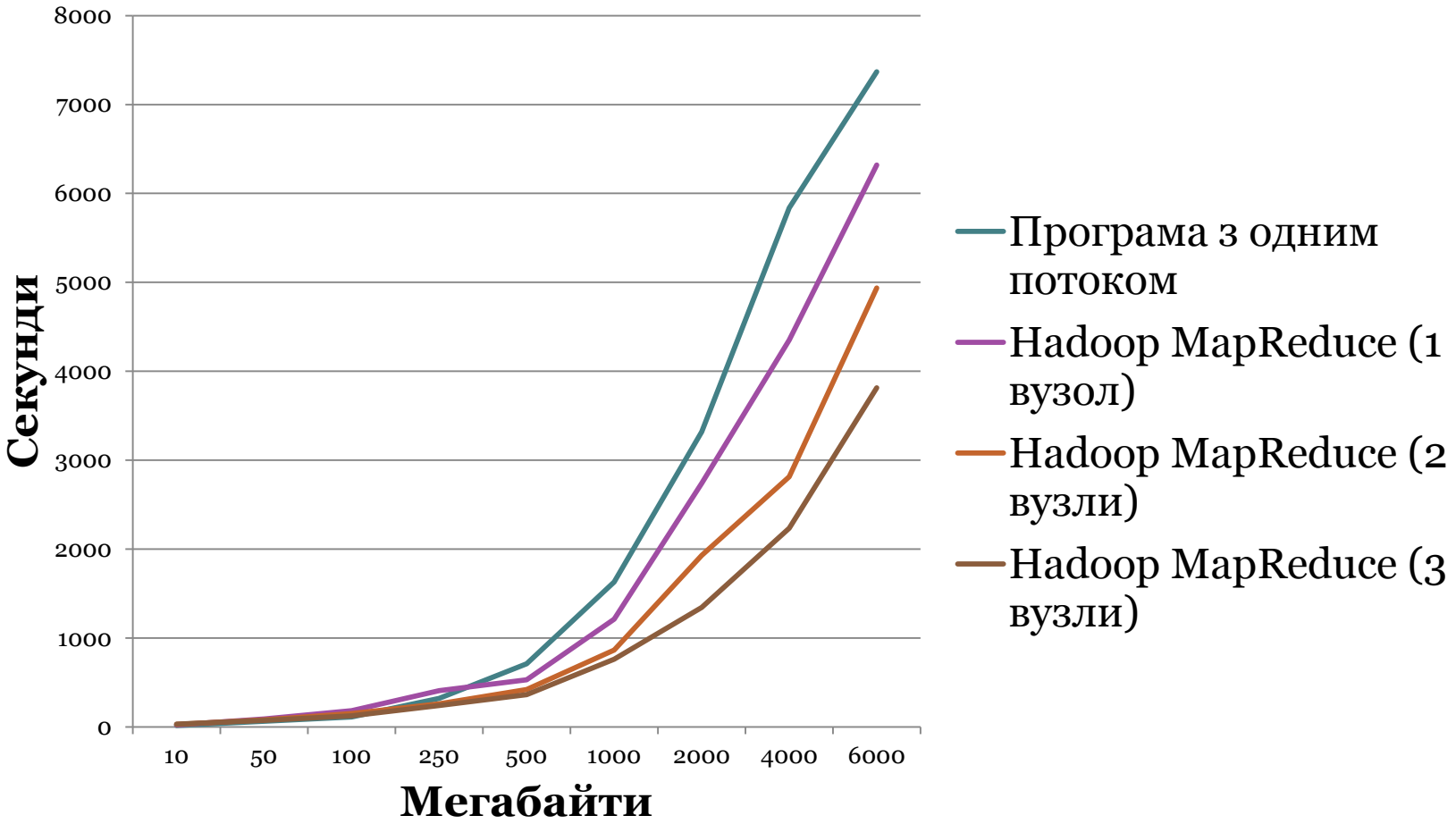




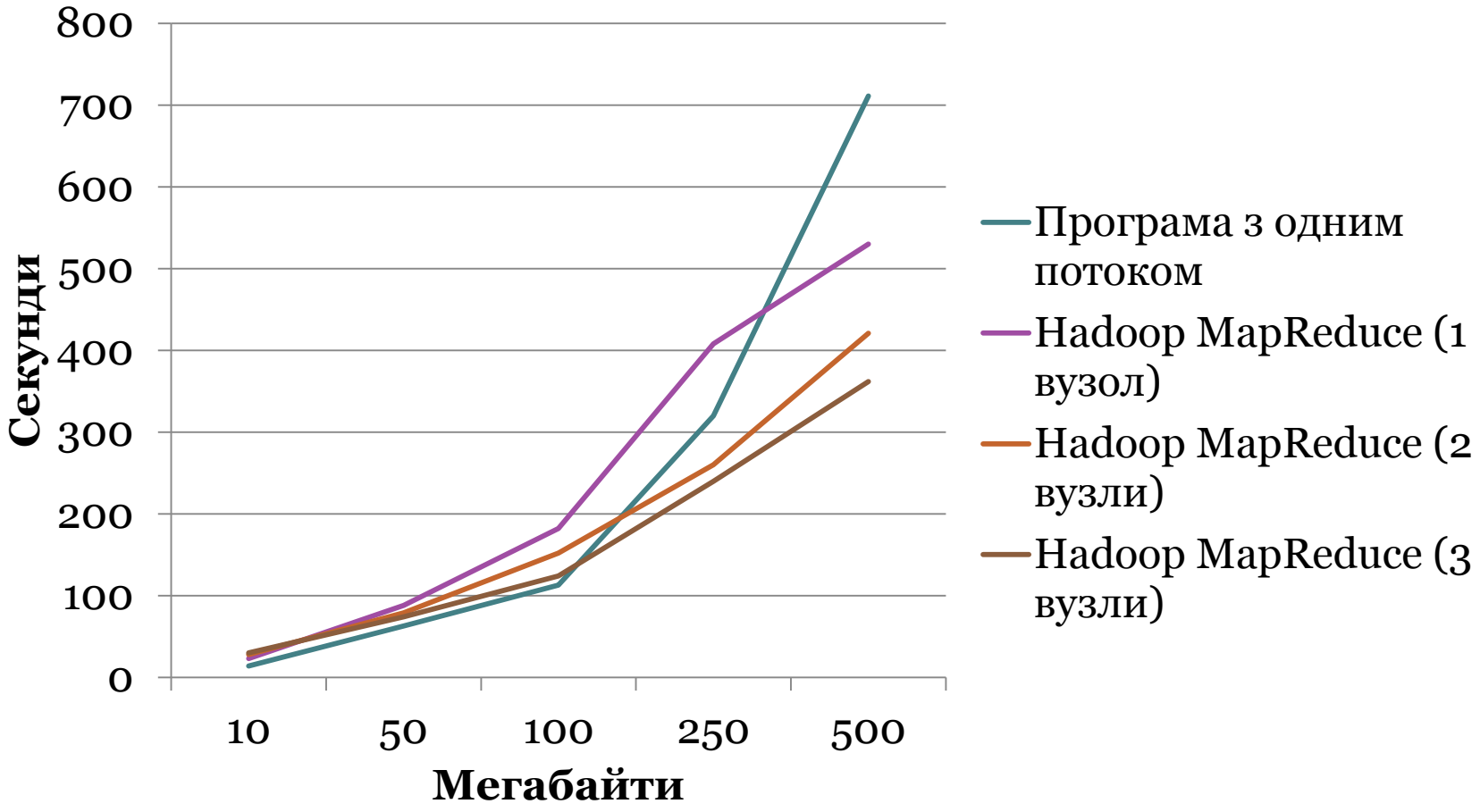




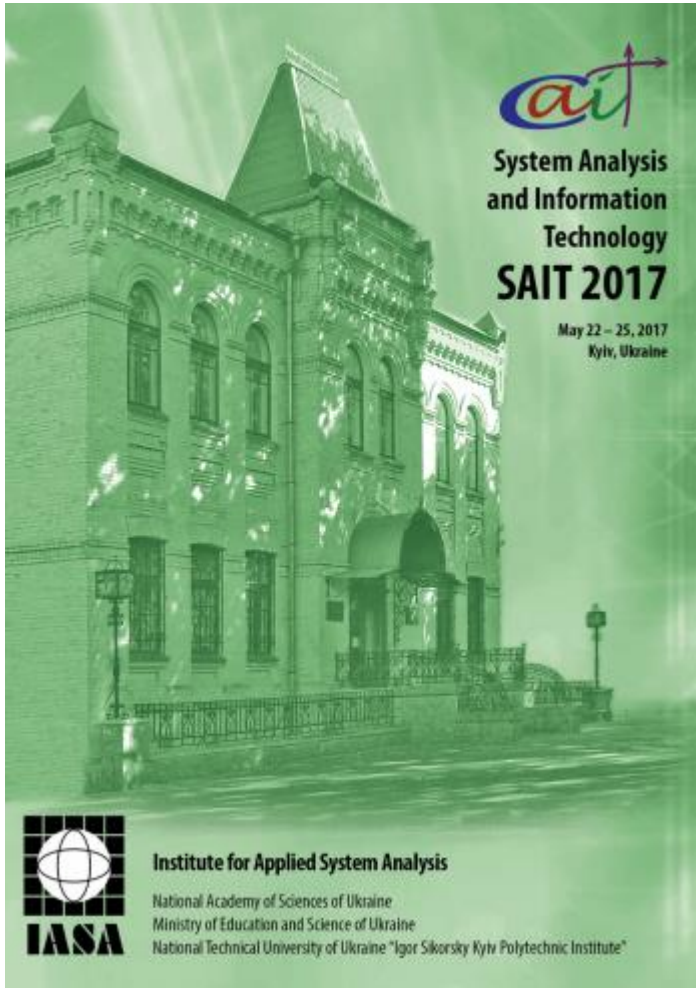
Аналіз швидкості роботи



Аналіз швидкості роботи (вхідні дані менші за 500 МБ)



Публікації



ISSN 2520-2057

МІЖНАРОДНИЙ
НАУКОВИЙ ЖУРНАЛ
«ІНТЕРНАУКА»

INTERNATIONAL
SCIETIFIC JOURNAL
«INTERNAUKA»

МЕЖДУНАРОДНЫЙ
НАУЧНЫЙ ЖУРНАЛ
«ИНТЕРНАУКА»

№ 6 (28) / 2017

Майбутні напрямки роботи та досліджень

- Викласти вихідні коди 

- Дослідити роботу моделі для вирішення задачі кластеризації в залежності від розподілення вхідних даних між вузлами
- Дослідити технологію 

- Дослідити поточний стан розвитку алгоритмів обробки великих даних з використанням квантових комп’ютерів

Дякую за увагу!

Студент групи ДА-52м
Яременко Вадим Сергійович