АВТОМАТИЗАЦІЯ ПРОЦЕСУ ПОШУКУ ГЕНІВ У ГЕНОМІ ЛЮДИНИ

Розробив: М.В. Плис

Керівник: С.М. Москвіна

Мета. Метою даної роботи є підвищення ефективності автоматизованої обробки та пошуку генетичної інформації в розподілених системах.

Об’єкт дослідження. Процеси пошуку слабоструктурованої генетичної інформації у розподіленому обчислювальному середовищі.

Предмет дослідження. Предметом дослідження є методи обробки та пошуку даних в великих масивах генома людини у розподіленому обчислювальному середовищі.

Наукова новизна. Розроблений метод пошуку генетичної даних як слабкоструктурованих даних великої розмірності, який на відміну від існуючих, використовує ієрархічне групування даних, що дозволяє оптимізувати дисковий простір та зменшити час пошуку.

ВИМОГИ ТА ОБМЕЖЕННЯ

- основний робочий формат геному – VCF (Variant Call Format) версій 41 та 42;
- використання розподіленого підходу до збереження генетичної інформації;
- можливість здійснення повнотекстового пошуку по множині файлів;
- можливість виконання аналітичних запитів з агрегатними функціями;
- обмеження по часу на виконання запитів – 5 секунд;
- кількість геномів – тисячі.

ЗРАЗОК VCF-ФАЙЛУ

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  NA19789
1   10616   rs376342519 CCGCCGTTGCAAAGGCGCGCCG  C   100 PASS    AC=4973;AF=0.993011;AN=5008;NS=25
1   14599   rs531646671 T   A   100 PASS    AC=739;AF=0.147564;AN=5008;NS=2504;DP=32081;EAS_AF=0.
1   14604   rs541940975 A   G   100 PASS    AC=739;AF=0.147564;AN=5008;NS=2504;DP=29231;EAS_AF=0.
1   14930   rs75454623  A   G   100 PASS    AC=2415;AF=0.482228;AN=5008;NS=2504;DP=42231;EAS_AF=0
1   15211   rs78601809  T   G   100 PASS    AC=3050;AF=0.609026;AN=5008;NS=2504;DP=32245;EAS_AF=0
1   15274   rs62636497  A   G,T 100 PASS    AC=1739,3210;AF=0.347244,0.640974;AN=5008;NS=2504;DP=
1   15644   rs564003018 G   A   100 PASS    AC=41;AF=0.0081869;AN=5008;NS=2504;DP=32966;EAS_AF=0.
1   15774   rs374029747 G   A   100 PASS    AC=60;AF=0.0119808;AN=5008;NS=2504;DP=22795;EAS_AF=0.
1   15820   rs2691315   G   T   100 PASS    AC=2056;AF=0.410543;AN=5008;NS=2504;DP=14933;EAS_AF=0
1   15903   rs557514207 G   GC  100 PASS    AC=2209;AF=0.441094;AN=5008;NS=2504;DP=7012;EAS_AF=0.
1   18849   rs533090414 C   G   100 PASS    AC=4767;AF=0.951877;AN=5008;NS=2504;DP=4700;EAS_AF=1;
1   30923   rs806731    G   T   100 PASS    AC=4369;AF=0.872404;AN=5008;NS=2504;DP=13565;EAS_AF=0
1   49298   rs200943160 T   C   100 PASS    AC=3917;AF=0.782149;AN=5008;NS=2504;DP=17078;EAS_AF=0
1   52238   rs2691277   T   G   100 PASS    AC=4616;AF=0.921725;AN=5008;NS=2504;DP=9078;EAS_AF=1;
1   54712   rs568927205 T   TTTTC   100 PASS    AC=2904;AF=0.579872;AN=5008;NS=2504;DP=20788;EAS_
1   55164   rs3091274   C   A   100 PASS    AC=4624;AF=0.923323;AN=5008;NS=2504;DP=12089;EAS_AF=1
1   55326   rs3107975   T   C   100 PASS    AC=230;AF=0.0459265;AN=5008;NS=2504;DP=16857;EAS_AF=0
1   55545   rs28396308  C   T   100 PASS    AC=1198;AF=0.239217;AN=5008;NS=2504;DP=15018;EAS_AF=0
1   57292   rs201418760 C   T   100 PASS    AC=169;AF=0.033746;AN=5008;NS=2504;DP=20190;EAS_AF=0.
1   62777   rs528401309 A   T   100 PASS    AC=1284;AF=0.25639;AN=5008;NS=2504;DP=20257;EAS_AF=0.
```

Представлення слабоструктурованих даних

Система містить дані з множини джерел $S = \{s_1, s_2, \dots, s_N\}$

Кожне джерело даних містить певну кількість записів $R_{S_k} = \{r_{k_1}, r_{k_2}, \dots, r_{k_m} \mid s_k \in S\}$.

Записи складаються з характеристик (полів) $P = \{p_1, p_2, \dots, p_L\}$

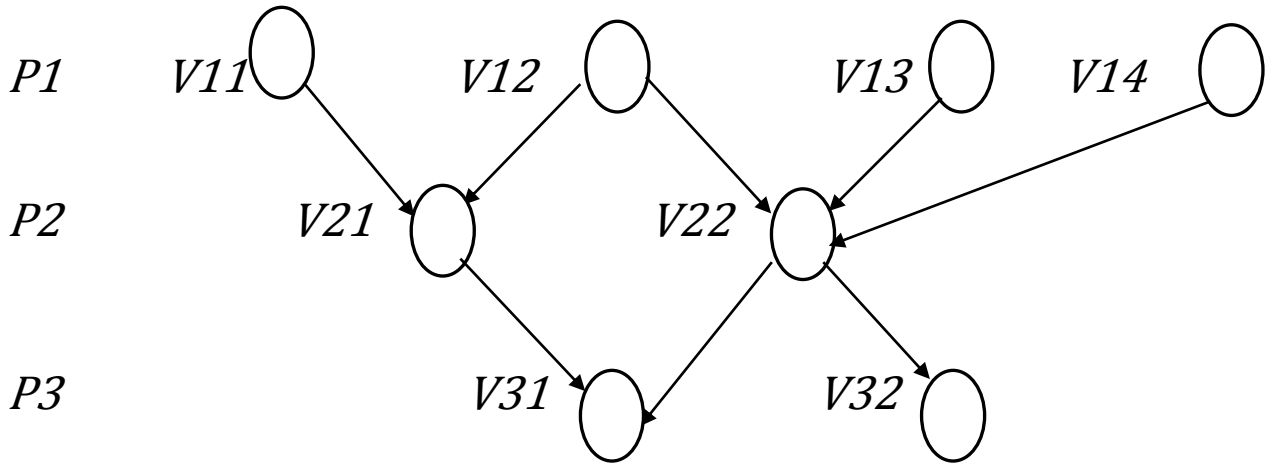
Унікальність запиту гарантується комбінацією підмножини його характеристик $P_d = \{p_{d_1}, p_{d_2}, \dots, p_{d_g} \mid p_d \in P\}$

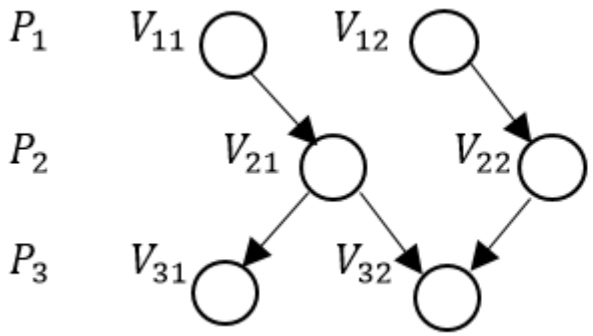
Задача аналізу даних: знайти всі джерела, в яких присутні задані характеристики $P_{dx} = \{s_1, s_2, \dots, s_t \mid s \in S\}$

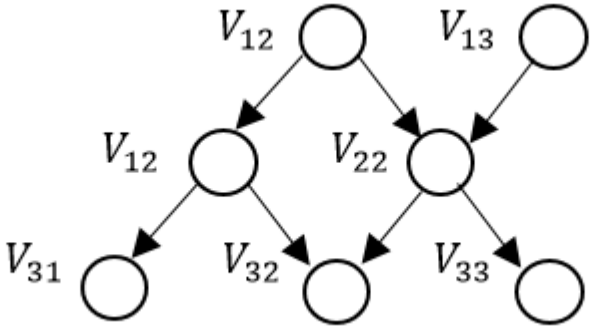
Модель зберігання слабоструктурованих даних

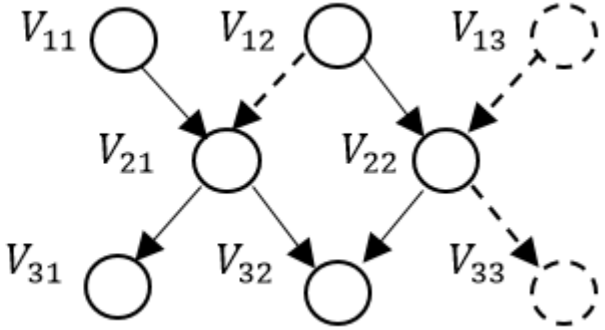
Дані описуються трьома характеристиками $\{P_1, P_2, P_3\}$, тому граф складається з трьох рівнів

Кожна характеристика (рівень графа) описується множиною свої значень. Так, наприклад, характеристика P_1 може приймати значення із множини $\{V_{11}, V_{12}, V_{13}, V_{14}\}$, P_2 з множини $\{V_{21}, V_{22}\}$, а P_3 з множини $\{V_{31}, V_{32}\}$.



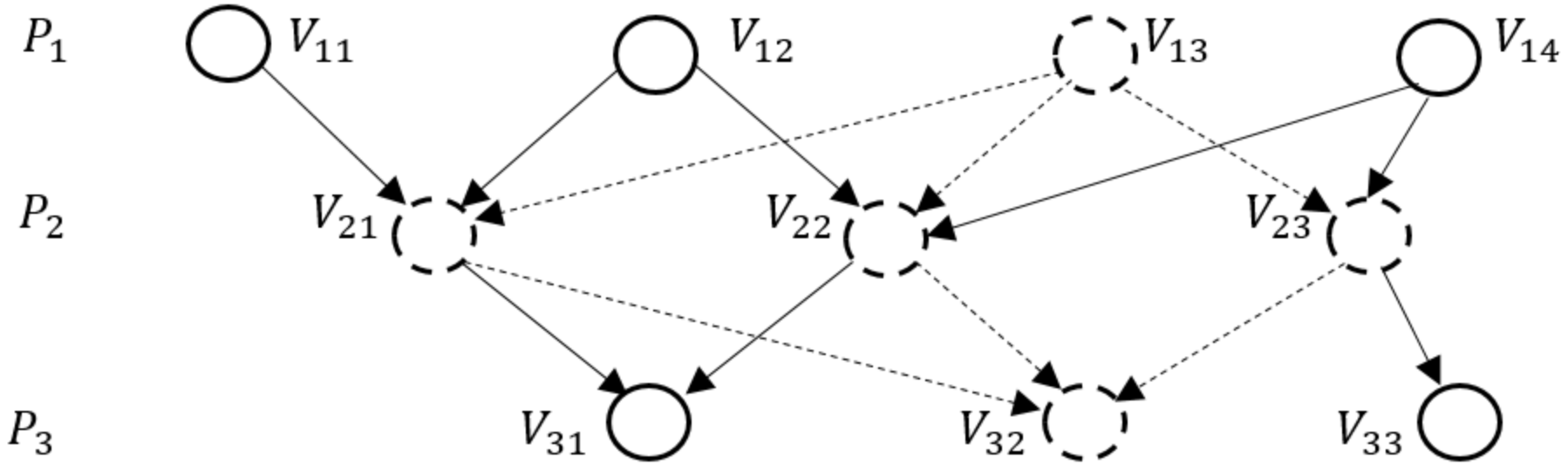






Реалізація пошуку з використанням розробленої моделі

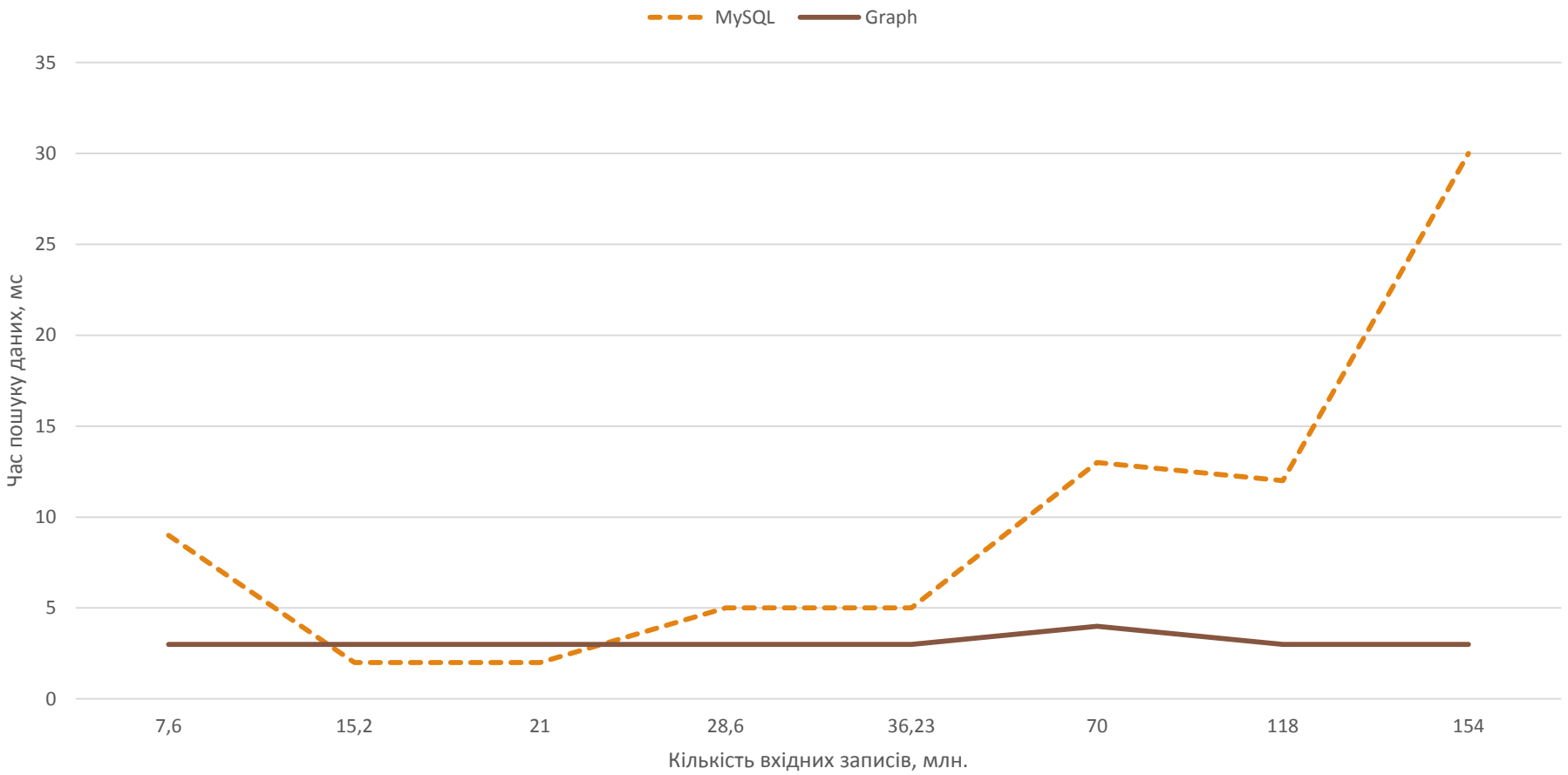
Припустимо, що є множина даних (модель якої показана на рисунку), і якщо пошук заданий множиною $\{V_{13}, V_{32}\}$, то будуть обрані об’єкти, які виділені від основного графа штрих-пунктирними лініями



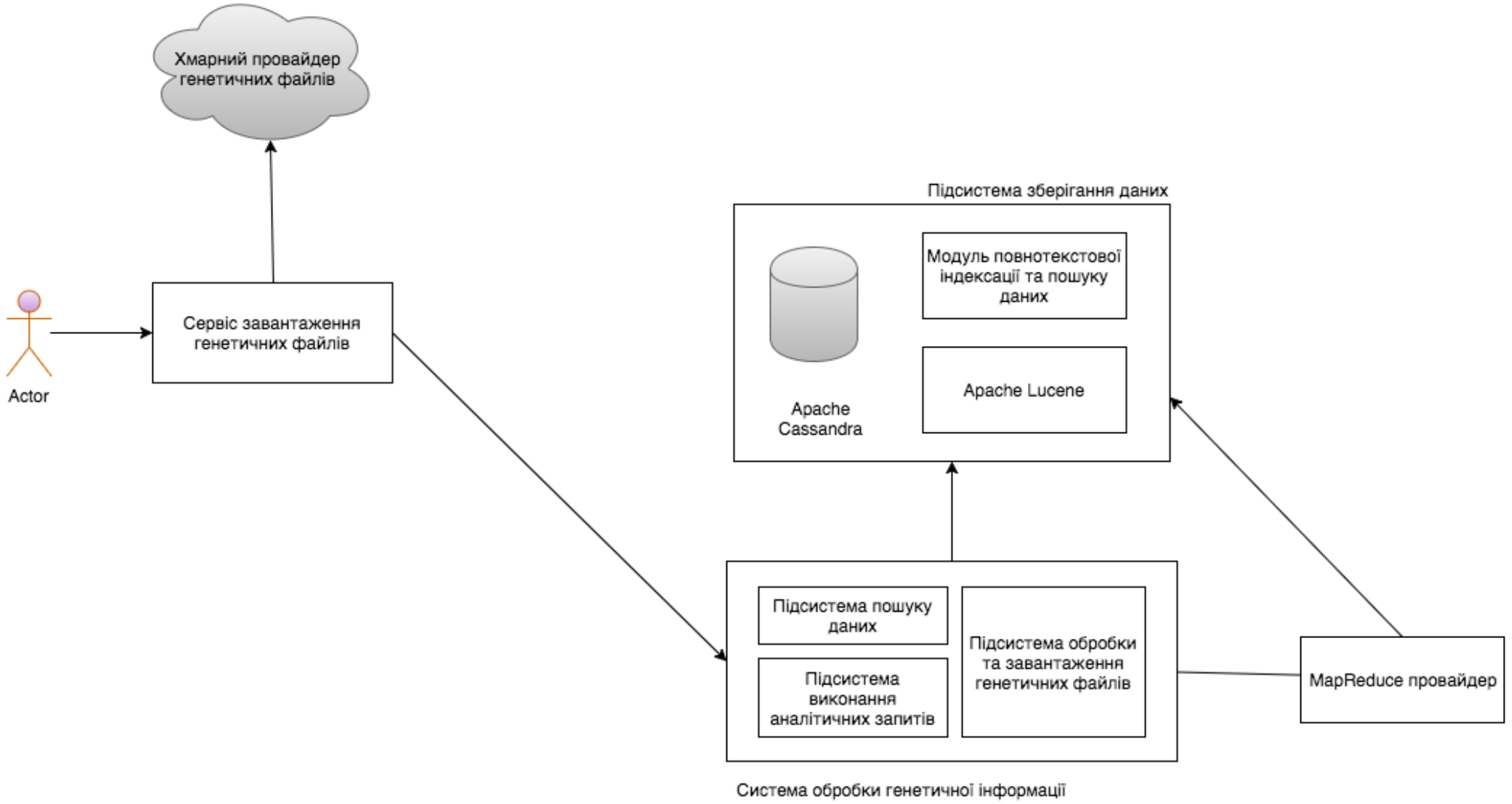
Експериментальні дослідження

Кількість завантаже-них файлів	Розмір файлів у файловій системі, Gb	Загальна кількість записів у файлах, млн.	Розмір бази даних, Gb		Час виконання запиту, мс	
			MySQL	Graph	MySQL	Graph
1	1,2	7,6	1,3	2	9	3
2	2,4	15,2	2,5	2,4	2	3
3	3,6	21	3,5	2,7	2	3
4	4,8	28,6	4,8	2,9	5	3
5	6	36,23	6,5	3,1	5	3
10	12	70	11,4	3,3	13	3
15	18	118	17	3,6	12	4
20	27	154	23	3,8	30	3

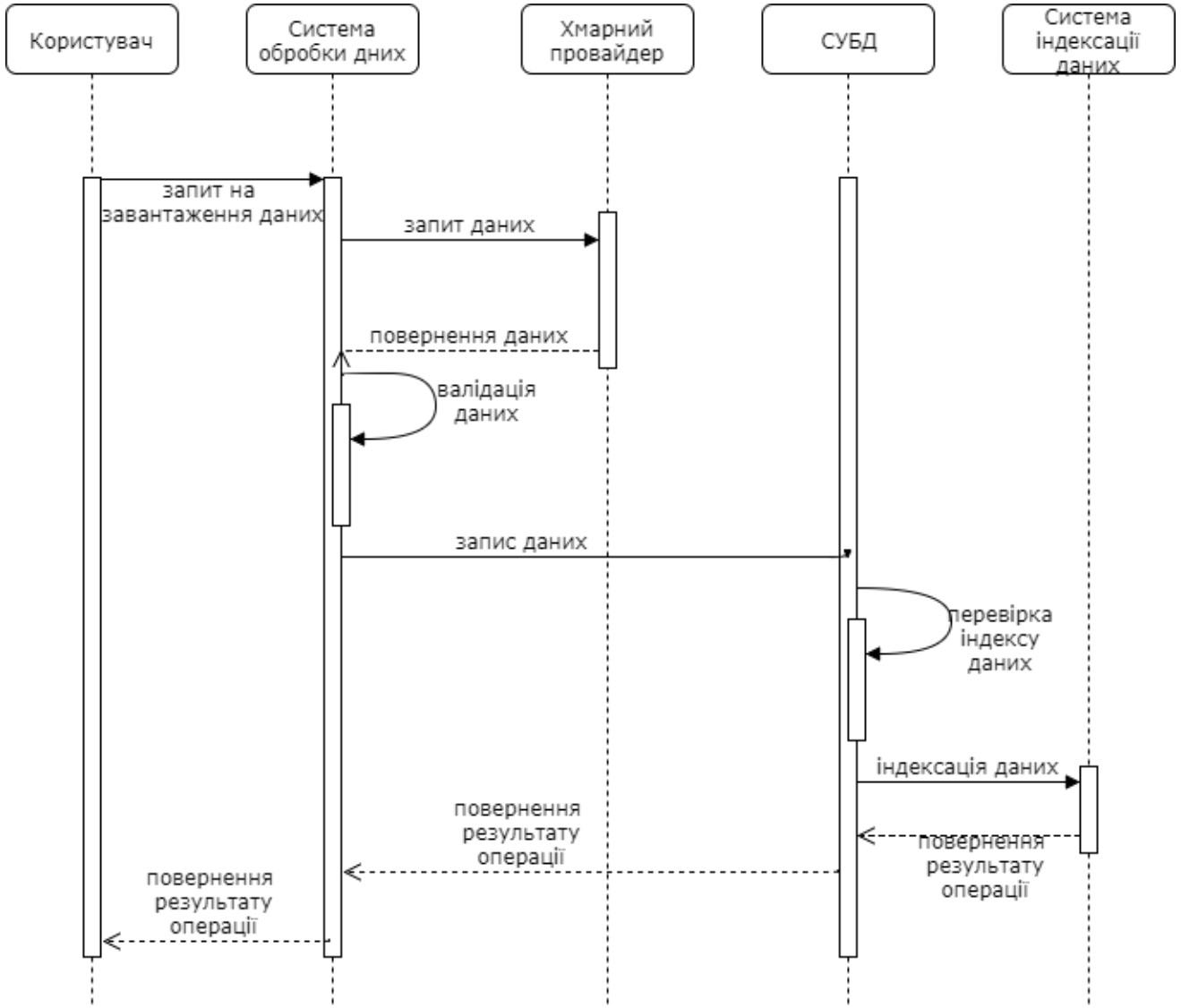
Експериментальні дослідження



Структурна схема розподіленої автоматизованої системи



UML-діаграма послідовності завантаження даних



Висновки

- Розроблений метод пошуку генетичних даних як слабкоструктурованих даних великої розмірності, який на відміну від існуючих, використовує ієрархічне групування даних, що дозволяє оптимізувати дисковий простір та зменшити час пошуку;

- Розроблена математична модель генетичної інформації як слабоструктурованих даних, в основі якої використовано графова ієрархічна структура даних, що дозволило оптимізувати розмір файлів бази даних;

- Розроблений алгоритм пошуку генетичної інформації як слабоструктурованих даних, оснований на використані розроблених в роботі операцій обробки слабоструктурованих даних, який дозволив суттєво зменшити час пошуку.

- Розроблена автоматизована системи пошуку генів у геномі людини без точок відмови із можливістю до масштабування відповідно до навантаження системи та можливістю подальшої модернізації для проведення конкретних обчислень над генетичними даними.

- Проведено дослідження та підтверджено доцільність використання розробленої автоматизованої системи пошуку генів у геномі людини

Публікація

- Докладено на XLV регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів ВНТУ та опубліковано тези доповідей;
- Докладено на XLVI регіональна науково-технічна конференція професорсько-викладацького складу, співробітників та студентів ВНТУ та опубліковано тези доповідей;
- Оформлено авторське свідоцтво на тему «Автоматизована система обробки генетичної інформації».