Выбор оптимальной базы данных для работы с социальным графом

И.В Гражданкин, email: grazhdankin.i.l@yandex.ru

С.В Власов , email: svv@cs.vsu.ru

Цель работы

- Цель работы – тестирование и анализ скорости выполнения запросов для реализаций графовых баз данных на примере социального графа.

Графы

Применение:

- Маршруты перевозки грузов
- Взаимосвязи пользователей социальных сетей

Особенности применения:

- Очень большой объем данных
- Требуется быстрый анализ графа

Реализации базы данных

Neo4j

- Самая распространённая база данных
- Обширный функционал
- Есть бесплатная версия

Sparksee

- Заявлена высокая производительность
- Есть бесплатная версия

Методология тестирования

Тестовый стенд:

- Intel Xeon X6550 2.0Gz
- 80Gb DDR3
- 2Tb hard drive

Настройка Neo4j:

- Version 3.5.11
- Ubuntu 18.04 LTS
- Cache size 60Gb

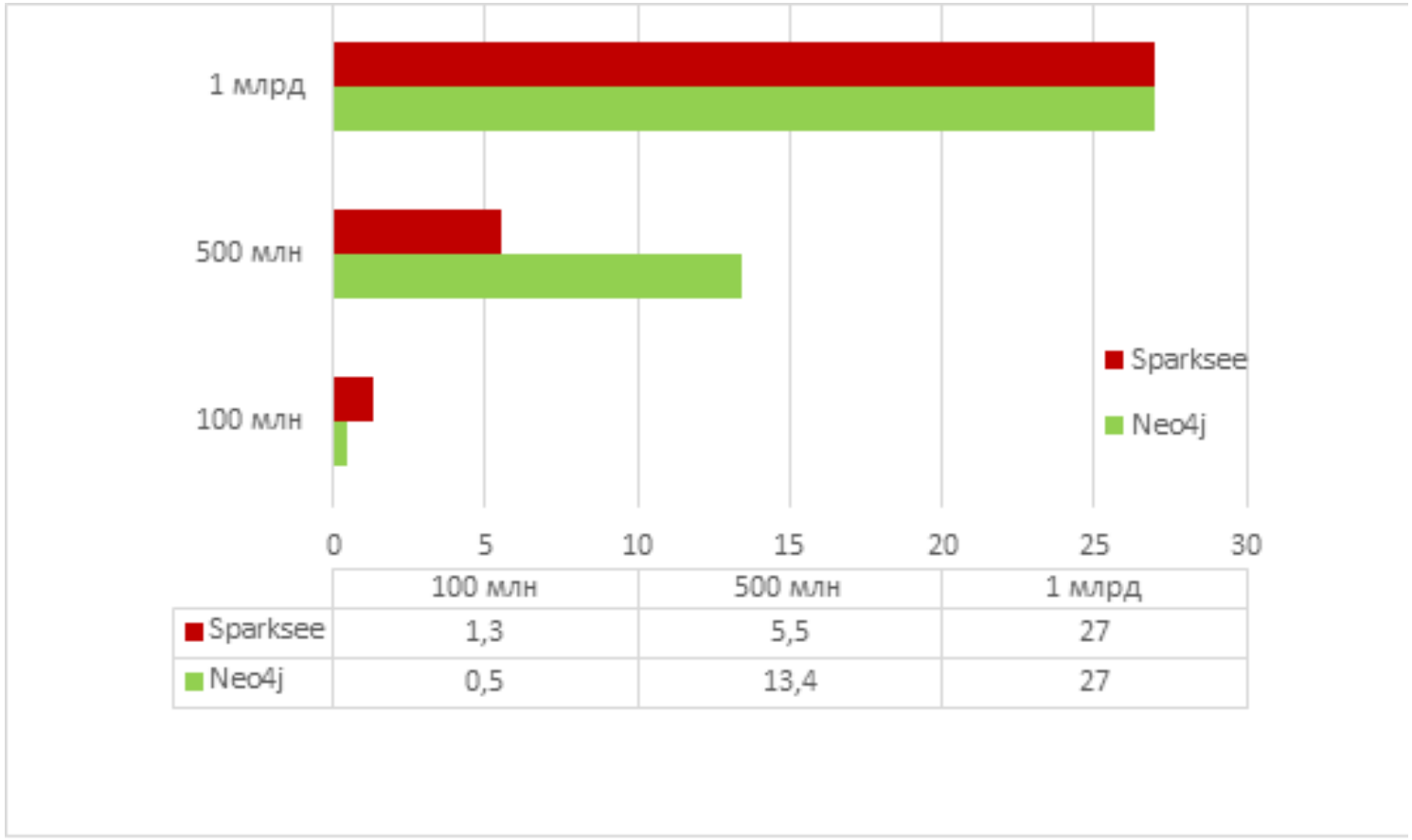
Настройка Sparksee:

- Version 5.2
- Ubuntu 18.04 LTS
- Cache size 60Gb

Аналитические запросы:

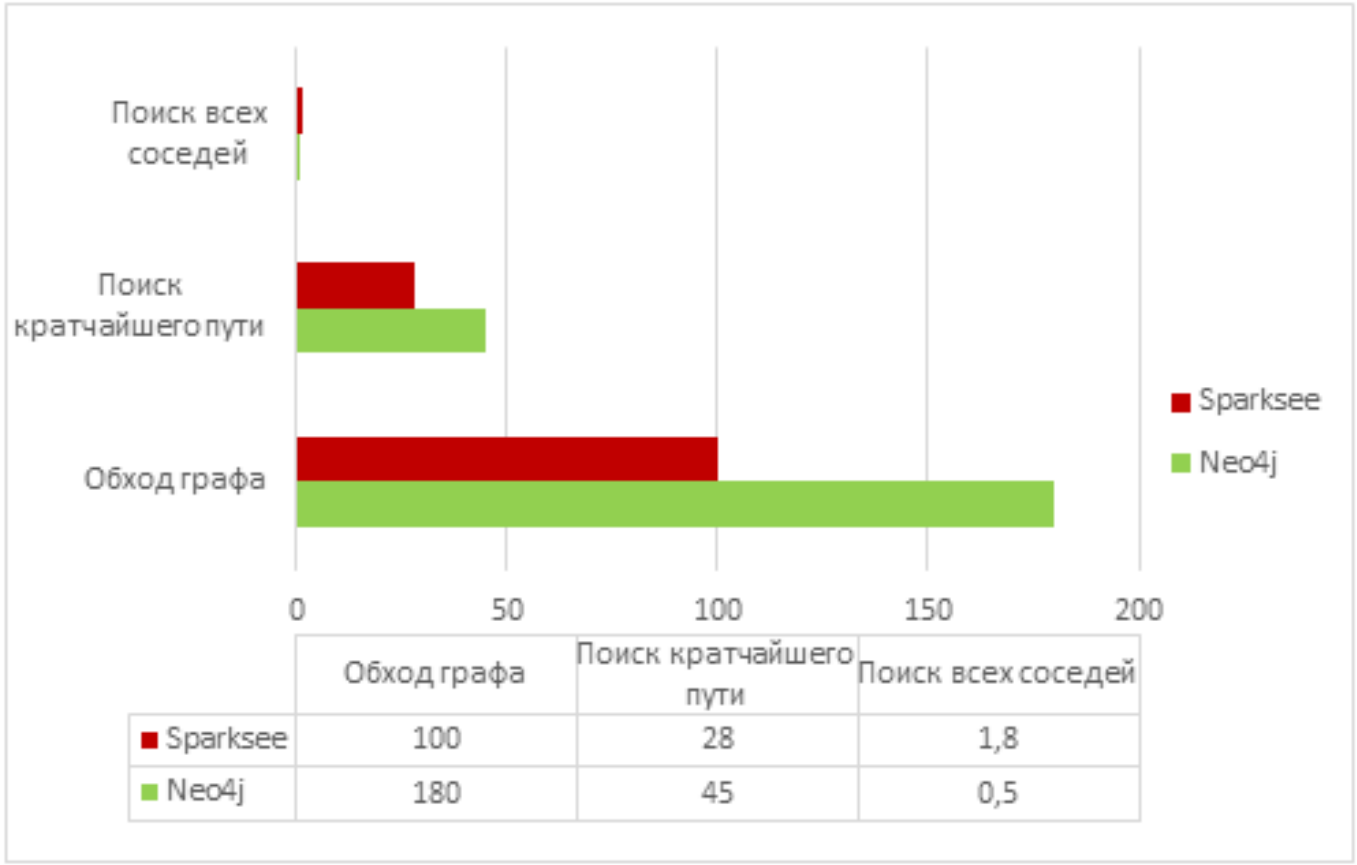
- Простой запрос – получить всех соседей вершины
- Более сложный запрос - найти кратчайший путь
- Сложный запрос – выполнить обход графа

Время импорта данных



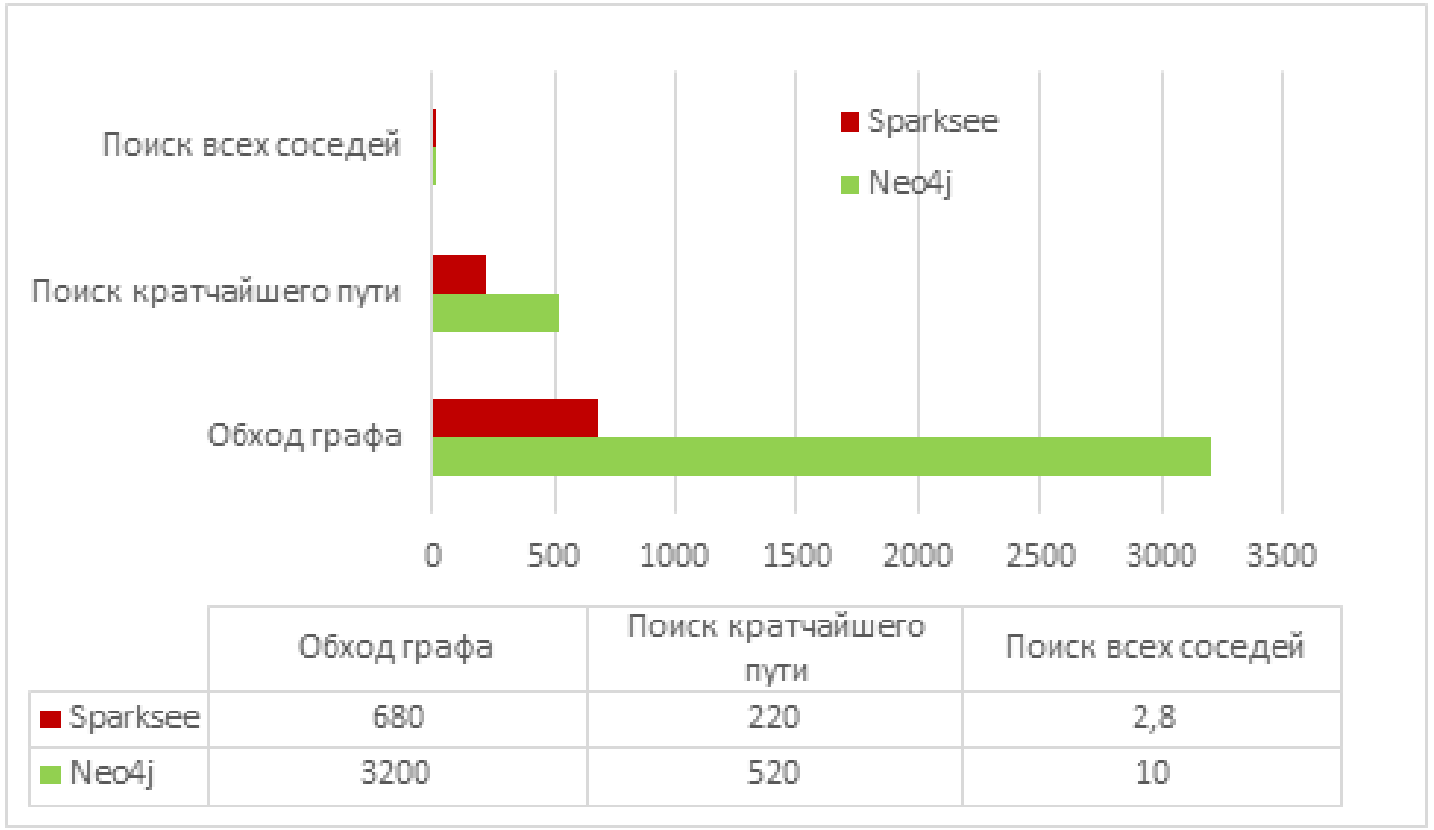
	100 млн	500 млн	1 млрд
Sparksee	1,3	5,5	27
Neo4j	0,5	13,4	27

Время обработки аналитических запросов 100 млн



	Обход графа	Поиск кратчайшего пути	Поиск всех соседей
Sparksee	100	28	1,8
Neo4j	180	45	0,5

Время обработки аналитических запросов 500 млн



	Обход графа	Поиск кратчайшего пути	Поиск всех соседей
Sparksee	680	220	2,8
Neo4j	3200	520	10

Выводы

- Время импорта занимает разумное время если, количество ребер не превышает одного миллиарда
- Sparksee предпочтительнее, чем Neo4j по производительности. Это очень хорошо видно при выполнении сложных аналитических запросов и при использовании большой части графа
- Главной характеристикой для высокой производительности графовых баз данных работы с графом, у которого количество ребер превышает один миллиард необходимо использовать кластерные решения