|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ Информатики и систем управления

КАФЕДРА Теоретической информатики и компьютерных технологий

**ОТЧЁТ ПО ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ**

Студент Александрова Ольга Сергеевна

*фамилия, имя, отчество*

Группа ИУ9-61Б

Тип практики Производственная

Название предприятия ЗАО «НОРСИ-ТРАНС»

Студент **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*подпись, дата фамилия, и.о.*

Рекомендуемая оценка: **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Руководитель практики   
от предприятия: **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*подпись, дата фамилия, и.о.*

Руководитель практики **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*подпись, дата фамилия, и.о.*

Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*2024 г*

Оглавление

[Введение 3](#_Toc173255696)

[1. Характеристика предприятия ЗАО «НОРСИ-ТРАНС» 4](#_Toc173255697)

[2. Индивидуальное задание 6](#_Toc173255698)

[2.1 Постановка задачи 6](#_Toc173255699)

[2.2 Решение задач 6](#_Toc173255701)

[2.2.1 Ознакомление с входными данными 6](#_Toc173255702)

[2.2.2 Создание docker контейнеров для баз данных 8](#_Toc173255703)

[2.2.3 Измерение количества памяти, занимаемых данными в базах данных 9](#_Toc173255704)

[2.2.4 Изучение и написание запросов 13](#_Toc173255705)

[2.2.5 Измерение времени и памяти, затраченных для каждого запроса 17](#_Toc173255707)

[Заключение 24](#_Toc173255708)

[СПИСОК ЛИТЕРАТУРЫ 25](#_Toc173255709)

# 

# Введение

Целью производственной практики является закрепление знаний по изучаемым дисциплинам и получение студентами практических навыков в период пребывания на предприятии, в связи с чем можно выделить следующие задачи данной практики:

1. Ознакомление с предприятием, его программными продуктами и предметной областью, для которой предназначены эти продукты.
2. Изучение жизненного цикла программного обеспечения, создаваемого на предприятии.
3. Изучение профессиональных взаимоотношений в рамках организационной структуры подразделения прохождения практики, организационных мер, направленных на эффективную совместную работу программистов и других специалистов над общим проектом.
4. Изучение процесса разработки программного обеспечения: какие модели разработки используются на предприятии, причины использования именно этих моделей, организационные меры и технические средства предприятия, направленные на поддержку этих моделей.
5. Получение навыков работы с промышленными средствами: автоматизированной сборки, отладки, контроля версий и разрешения конфликтов версий, анализа кода, автоматизированного тестирования, профилирования.
6. Обобщение полученных результатов и навыков, составление отчёта по практике.

# 

# 1. Характеристика предприятия ЗАО «НОРСИ-ТРАНС»

ЗАО «НОРСИ-ТРАНС» – российский разработчик и производитель серверного оборудования, систем хранения данных на различных платформах, телекоммуникационного оборудования и вычислительных платформ.

ЗАО «НОРСИ-ТРАНС» работает с 1995 года и специализируется на создании и выпуске:

* серверного оборудования;
* высокоплотных систем хранения данных;
* HPC-вычислительных платформ;
* телекоммуникационного оборудования;
* высокосложных корпусов включая полный цикл выпуска КД;
* печатных плат в составе выпускаемой продукции;
* системного и прикладного программного обеспечения для различных аппаратных платформ.

Компания разрабатывает и создает готовые программно-аппаратные комплексы под специализированные требования заказчиков. Клиенты Компании - операторы сотовой подвижной связи, операторы фиксированной связи, интернет - провайдеры. ЗАО "НОРСИ-ТРАНС" ведет сотрудничество c крупнейшими российскими поставщиками услуг связи, такими как «МТС», «Мегафон», «Ростелеком», «Национальные кабельные сети», «Нэт Бай Нэт Холдинг».

Кроме этого, в компании ведутся уникальные разработки с использованием собственных программных продуктов, научно-исследовательские и опытно-конструкторские работы. ЗАО «НОРСИ-ТРАНС» с мая 2009 года входит в состав общественно-государственного объединения «Ассоциация документальной электросвязи» и с 2010 года входит в состав наблюдателей, а с 2015 года - членом Европейского Института Телекоммуникационных Стандартов (ETSI).

В этом году, для студентов провели лекции, на которых представители направлений компании рассказывали о задачах, базовых концепциях и продуктах Компании. Лекции затрагивали следующие темы:

1. Введение в архитектуры нейронных сетей. Языковые модели (NLP, Natural Language Processing).
2. Методы поиска аномалий в данных.
3. Распознавание текста и его коррекция.
4. Рекуррентные нейронные сети.
5. Инструмент проектирования Web-интерфейса, разработанный в ЗАО «НОРСИ-ТРАНС»
6. Многообразие современных СУБД.
7. СУБД ClickHouse.
8. Загрузка данных в хранилища данных (ETL - Extract / Transform / Load).
9. Архитектура и принцип работы криптовалюты Bitcoin.
10. Криптовалюта ETHEREUM.
11. Опыт курсового и дипломного проектирования.

# 2. Индивидуальное задание

## 2.1 Постановка задачи

В рамках практики была поставлена цель - сравнение производительности баз данных, таких как ArangoDB, Dgraph и Scylla. Индивидуальное задание состояло из следующих задач:

* ознакомиться с предложенными датасетами;
* создать docker контейнер для каждой из баз данных;
* изучить объем памяти, занимаемый данными в базах;
* изучить и написать запросы для каждого графа;
* измерить время и память, затраченные для каждого запроса;
* проанализировать результаты измерений и сделать выводы.

## Выбор подходящей базы данных для хранения и обработки данных является основным фактором при разработке приложений и систем. Существует множество различных баз данных, каждая из которых имеет свои особенности и преимущества. Правильный выбор базы данных поможет улучшить обработку запросов, снизить нагрузку на серверы. В рамках практики стоит цель - сравнение производительности трех популярных баз данных: ArangoDB, Dgraph, Scylla.

## 2.2 Решение задач

**2.2.1 Ознакомление с входными данными**

Оценка производительности баз данных рассчитывается при выполнении запросов над следующими датасетами:

1) MOOC User Action Dataset [2];

2) Elliptic++ Transactions Dataset [3];

3) California road network [4];

4) Stablecoin ERC20 Transactions Dataset [5].

Набор данных о действиях пользователей MOOC представляет действия, предпринятые пользователями на популярной платформе MOOC. Действия представлены в виде направленного временного графа. Узлы представляют пользователей и действия курса (цели), а ребра представляют действия пользователей на целях. Действия имеют атрибуты и временные метки. Размер исходных файлов: mooc\_action\_features.tsv - 35,9 МБ, mooc\_action\_labels.tsv - 3,6 МБ, mooc\_actions.tsv - 10,6 МБ. В итоге получаем 50.1 МБ.

Набор данных транзакций Elliptic++ состоит из 203 тыс. транзакций биткоин. Каждая транзакция имеет 183 атрибута, однако для упрощения выгрузки данных было решено взять 19 полей, а именно: 'txId', 'time\_step', 'Local\_feature\_1','Aggregate\_feature\_1', 'in\_txs\_degree', 'out\_txs\_degree', 'total\_BTC', 'fees', 'size', 'num\_input\_addresses', 'num\_output\_addresses', 'in\_BTC\_min', 'in\_BTC\_max', 'in\_BTC\_mean', 'in\_BTC\_median', 'in\_BTC\_total', 'out\_BTC\_min', 'out\_BTC\_max', 'out\_BTC\_mean', 'out\_BTC\_median', 'out\_BTC\_total'.

Размер исходных файлов: txs\_edgelist.csv - 4,5 МБ, txs\_classes.csv - 2,4 МБ, txs\_features.csv - 44,3 МБ (из таблицы были удалены неиспользуемые атрибуты). В итоге получаем 51.2 МБ.

California road network представляет собой дорожную сеть штата Калифорния, построенную с использованием графовой модели. Узлы - представляют собой пересечения дорог или конечные точки дорог. Ребра - представляют собой дороги, соединяющие узлы. Ребра являются ненаправленными, что означает, что движение по ним возможно в обоих направлениях. Размер исходного файла roadNetCA.txt - 87,8 МБ.

Stablecoin ERC20 Transactions Dataset содержит данные о транзакциях ERC20 валюты по рыночной капитализации (USDT, USDC, DAI, UST, PAX) и WLUNA. Каждая транзакция в данных включает в себя количество соответствующей валюты, определенной адресом контракта в строке, передаваемых между различными внешними адресами. Набор данных содержит более 70 миллионов транзакций, включает в себя данные о ценах и содержит данные о событиях (событие, временная метка, тип, валюта). Для наших целей было принято взять файл token\_transfers\_V2, выбрав из него для импорта те транзакции, у которых contract\_address равен 0xdac17f958d2ee523a2206206994597c13d831ec7. В итоге файл token.csv занимает 2,5 ГБ.

**2.2.2 Создание docker контейнеров для баз данных**

Создание Docker контейнеров для баз данных, таких как ArangoDB, Dgraph и Scylla, предоставляет ряд преимуществ, которые делают развертывание и использование этих баз данных более удобным и эффективным по данным причинам:

1. Портативность: Docker контейнеры являются переносимыми и могут быть развернуты на любой совместимой системе. Это означает, что есть возможность создать контейнер на одной машине и легко перенести его на другую без необходимости повторной установки и настройки базы данных.

2. Изолированность: Каждый Docker контейнер работает в изолированной среде, что означает, что база данных и все ее зависимости находятся внутри контейнера и не влияют на другие процессы или системы. Это обеспечивает стабильность и безопасность работы базы данных.

3. Масштабируемость: Docker контейнеры позволяют легко масштабировать базу данных в зависимости от потребностей проекта. Есть возможность запускать несколько контейнеров одновременно или масштабировать контейнеры горизонтально для обработки больших объемов данных и запросов.

В ходе работы были успешно созданы контейнеры для каждой базы данных и выгружены, перенесены данные с датасетами.

**2.2.3 Измерение объема памяти, занимаемых данными в базах данных**

Таблица 1 представляет информацию о размерах датасетов и объеме памяти, занимаемым в базе данных. Эта таблица позволяет оценить и сравнить объем памяти, необходимый для их хранения и обработки.

Таблица 1 – Занимаемая память в датасетах

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **База данных** | **Dgraph** | **ArangoDb** | **Scylla** | **Исходные данные** |
| **Mooc** | 49.50563 MB | Ребра  размер документов: 17.04 МБ  размер индексов: 38.99 МБ  Вершины  размер документов: 130.81 kB  размер индексов: 94.14 kB  в сумме **56.25 МБ.** | Space (total): 38,06112 МБ  Bloom filter space: 2.433540 MB  Index summary: 42.890625 КБ | 50.1 MB |
| **Elliptic** | 37.6 МВ. | Ребра  размер документов: 24.68 MB  размер индексов: 33.91 MB  Вершины  размер документов: 103.21 MB  размер индексов: 9.98 MB  в сумме **171.78 МБ.** | Space (total): 33,7 MB  Bloom filter space: 1,3 МB  Index summary: 46,296 КB | 51.2 MB. |
| **RoadNet** | 2,342.758245 MB. | Ребра  doc размер документов: 115.99 MB  размер индексов: 541.67 MB  Вершины  размер документов: 35.12 MB  размер индексов: 27.37 MB  в сумме **720.15 МБ.** | Space (total): 94.963663 МБ  Bloom filter space: 3.20986 MB  Index summary: 0.103176 MB | 87,8 MB |
| **ERC20** | 373.2 MB | Ребра  размер документов: 1.37 GB  размер индексов: 2.11 GB  Вершины  размер документов: 217.96 MB  размер индексов: 216.86 MB  в сумме **3.90 ГБ** | Space (total): 1.804976 GB  Bloom filter space: 17.50002 MB  Index summary: 2.43436 MB | 2,5 ГБ |

Анализ таблицы позволяет сделать следующие выводы:

• ArangoDB потребляет больше памяти, чем Dgraph и Scylla во всех датасетах, кроме RoadNet. Это связано с тем, что ArangoDB использует графовую модель данных, которая может занимать достаточно много места для хранения связей между узлами.

• Разница в хранении данных в памяти между базами данных заметно увеличивается с ростом размера датасета.

• RoadNet в Dgraph занимает существенно больше памяти, чем ArangoDB и Scylla.

• Для задач, где требуется высокая производительность и размер датасета не критичен, ArangoDB может быть хорошим выбором.

Также можно видеть - ArangoDB предоставляет информацию о размере документов и индексов. Scylla дает представление о различных компонентах, занимающих память в базе данных (объема памяти для Bloom фильтра, индексной сводки). Dgraph не предоставляет детальной информации о компонентах, занимающих память.

Таблица 2 представляет информацию о количестве вершин и ребер в графе.

Данные Elliptic в Dgraph представлены таким образом: вершины — транзакции, обладающие предикатами, которые характеризуют транзакции: class, Local\_feature\_1 и остальными атрибутами, которые были перечислены в ознакомлении с входными данными. Связь с другими вершинами, следующими относительно данной транзакции, выражается через предикат successors.

MOOC в Dgraph: вершины — пользователи платформы, курсы на платформе и взаимодействия пользователей с этими курсами. Предикат performs определяет действия, совершённые данным пользователем. Предикат on определят курс, с которым связано пользовательское действие. Действия также обладают предикатами, характеризующими их атрибуты: label, feature0, feature1, feature2, feature3, timestamp.

Данные RoadNet в Dgraph: вершины — пункты назначения или пересечения дорог. Вершины обладают предикатом successors (ребро), определяющим вершины, непосредственно достижимые из данной.

ERC20 в Dgraph: вершины — внешние адреса, с которыми ассоциированы транзакции, и отправления, совершённые в рамках транзакций. Адрес обладает предикатом from, определяющим отправления от данного адреса. Отправление обладает предикатом to, адрес назначения, и предикатами, характеризующими это отправление: block\_number, transaction\_index, time\_stamp, contract, value .

Elliptic в ArangoDB был выгружен данным образом: коллекция вершин, первичным ключом которых является идентификатор транзакции, и каждая транзакция (вершина) хранит такие же атрибуты, как и в Dgraph, которые раннее были перечислены. Каждое ребро содержит в себе два первичных ключа двух вершин, которые имеют связь.

Логика представления данных MOOC в ArangoDB совпадает с логикой представления в Dgraph, атрибуты каждого ребра также совпадают.

RoadNet в ArangoDB представлен двумя документами – вершины и ребра. Вершины хранят в себе идентификатор узла, а в каждом ребре два ключа вершин, которые имеют связь.

Представление датасета ERC20 в ArangoDB такое же, как и в MOOC, за исключением другого набора атрибутов у каждого ребра - time\_stamp, value. Так как поле contract имеет одно и то же значение, оно не было выгружено.

Scylla не является графовой базой данной – данные напрямую записывались из таблицы в таблицу со всеми атрибутами, поэтому она будет представлять из себя количество строк из таблиц связей каждого графа.

Таблица 2 – Количество вершин и ребер

|  |  |  |  |
| --- | --- | --- | --- |
| **База данных** | **Dgraph** | **ArangoDb** | **Scylla** |
| **Кол-во вершин**  **Mooc** | - user: 7047  - action: 411749  - target: 97  Сумма: 418893 | - user: 7047  - target: 97  Сумма: 7144 | **-** |
| **Кол-во ребер**  **Mooc** | Рёбра:  -performs: 411749  - on: 411749  Сумма: 823498 | 411749 | 411748 |
| **Кол-во вершин**  **Elliptic** | 203769 | 203769 | - |
| **Кол-во ребер**  **Elliptic** | (без учёта обратных): 234355 | 234355 | 234355 |
| **Кол-во вершин**  **RoadNet** | 1965206 | 1965206 | - |
| **Кол-во ребер**  **RoadNet** | (с учётом обратных): 5533214 | 5533214 | 1965206 |
| **Кол-во вершин**  **ERC20** | - address: 1523333  - transfer: 5280131  Сумма: 6803464 | 3671592 | - |
| **Кол-во ребер**  **ERC20** | - from: 5280131  - to: 5280131  Сумма: 10560262 | 14650275 | 14443083 |

**2.2.4 Разработка запросов**

Ниже представлено описание каждого запроса, после каждого – таблица с параметрами входных данных для анализа. Так как у RoadNet нет атрибутов – второй и четвертый запрос невозможен, а в третьем не используем фильтрацию по полю.

После каждого описания запроса представлена таблица с значениями, которые использовались для тестирования.

1) Первый запрос предназначен для поиска вершин графа, которые связаны с заданной стартовой вершиной на заданной глубине, и имеют значение определенного параметра, большее или равное заданному значению. Обход совершается в глубину и в ширину.

Стоит сказать, что в Dgraph нет полноценного DFS или BFS, но есть рекурсивные запросы, которые имитируют BFS.

Таблица 3 – Параметры первого запроса

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр входных данных | Стартовая вершина (идентификатор) | Глубина | Поле для фильтрации | Значение для сравнения |
| **Mooc** | User/6725 | 1 | timestamp | 2571264.000 |
| **Elliptic** | 158652955 | 9 | size | 100 |
| **RoadNet** | 1086292 | 15 | - | - |
| **ERC20** | 0x4f2c813dc8078a9ea7884ff287c0176603e7d1db | 8 | value | 493112 |

1. Стартовая вершина - идентификатор начальной вершины, от которой будет производиться поиск.

2. Глубина - количество ребер, которое необходимо пройти от начальной вершины (startVertex).

3. Поле для фильтрации - параметр, по которому будет производиться фильтрация.

4. Значение для сравнения - значение, по которому будут отбираться ребра.

Далее описание повторяющих полей в последующих запросах совпадает и будет пропущено.

2) Второй запрос выполняет фильтрацию всех вершин по указанному полю и значению, возвращая список словарей с идентификаторами и значениями этого поля. Данный запрос получает множество вершин, по каждой из которых происходит перебор.

Таблица 4 – Параметры второго запроса

|  |  |  |
| --- | --- | --- |
| Параметр входных данных | Поле для фильтрации | Значение для сравнения |
| **Mooc** | timestamp | 2571264.000 |
| **Elliptic** | size | 85832 |
| **RoadNet** | - | - |
| **ERC20** | value | 101000000 |

3) Третий запрос выбирает вершины по значению поля и фильтрует по количеству, которое должно превышать определенное количество, исходящих ребер из соответствующей вершины. Он, как и во втором запросе, проходит по каждой вершине, при этом считая количество смежных (исходящих) ребер - степень вершины.

Таблица 5 – Параметры третьего запроса

|  |  |  |  |
| --- | --- | --- | --- |
| Параметр входных данных | Поле для фильтрации | Значение для сравнения | Степень вершины для сравнения |
| **Mooc** | timestamp | 2571264.000 | 2 |
| **Elliptic** | size | 85832 | 12 |
| **RoadNet** | - | - | 10 |
| **ERC20** | value | 493112.608952 | 5 |

Степень вершины для сравнения - минимальная степень вершины, которая будет учитываться в фильтрации.

4) Четвертый запрос выполняет агрегацию по удовлетворяющим полям соседних вершин. Идея данного запроса заключаются в том, чтобы вычислить общую сумму значений определенного поля (поле для фильтрации), значение которого должно превышать определенное число, для всех смежных ребер каждой вершины в заданном документе или таблицы. В запросе перебираются все вершины в заданном на входе множестве.

Таблица 6 – Параметры четвертого запроса

|  |  |  |
| --- | --- | --- |
| Параметр входных данных | Поле для фильтрации | Значение для сравнения |
| **Mooc** | timestamp | 2571264.000 |
| **Elliptic** | size | 300 |
| **RoadNet** | - | - |
| **ERC20** | value | 300000 |

5) Пятый запрос предназначен для поиска кратчайшего пути между двумя вершинами в графе. Данный запрос принимает на вход начальную вершину и конечную вершину. Результатом запроса является список идентификаторов вершин и ребер, образующих этот кратчайший путь. Этот запрос был реализован с помощью готового функционала в ArangoDB и Dgraph.

Пятый запрос не был реализован для базы данных Scylla, поэтому его значения будут отсутствовать в результатах, приведенных ниже в пункте 2.2.5.

Таблица 6 – Параметры пятого запроса

|  |  |  |
| --- | --- | --- |
| Параметр входных данных | Стартовая вершина | Целевая вершина |
| **Mooc** | User/1783 | Target/6 |
| **Elliptic** | 10853450 | 157642505 |
| **RoadNet** | 1518143 | 1517205 |
| **ERC20** | 0x5f407a10e031c0f  4b6c2286089023b01267bb01c | 0x2027a4a5884d7c511dfb  542d5c68650471564215 |

**2.2.5 Измерение времени и памяти, затраченных для каждого запроса**

В результате работы были запущены и проанализированы данные, полученные при запуске запросов, а именно – получено время и используемая память, которая была посчитана как разница между свободной оперативной памяти до и после выполнения запроса.

Для получения наиболее объективных результатов в ArangoDB для каждого запроса была отключена опция использования данных кэша.

Кэширование поддерживается в Dgraph, но только для запросов GraphQL (GraphQL — это ещё один язык запросов в Dgraph. Он обладает ещё меньшей выразительной силой в сравнении с DQL, поскольку предназначен для написания запросов к API от внешних непроверенных клиентов) и только на стороне браузера или CDN(Content Delivery Network). Кэширования на стороне сервера, соответственно, нет. Так как запросы были выполнены на языке запросов DQL результаты можно считать объективными.

Scylla использует кэширование запросов, но при перезапуске доккер контейнера кэши не сохраняются.

Оценка производительности была произведена на устройстве MacBook Air, Apple M1, память: 8 ГБ, общее количество ядер: 8 (4 производительности и 4 эффективности), загрузочный том: Macintosh HD, частота центрального процессора Apple M1 составляет 3,20 ГГц.

Таблица 7 демонстрирует результаты выполнения запросов к набору данных MOOC User Action, проведенных для сравнения производительности баз данных Dgraph, ArangoDB и Scylla.

Таблица 7 – Результаты для MOOC User Action Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Запросы\База данных** | **Dgraph** | **ArangoDb** | **Scylla** |
| **BFS** | 3872 Kb  0.009059 s | 29.000000 Kb 0.009387 s | 51.789062 Kb 0.644956 s |
| **DFS** |  | 29.000000 Kb 0.006746 s | 74.218750 Kb 0.523199 s |
| **Второй запрос** | 28960 Kb  0.00482 s | 10.911133 Kb 0.035456 s | 4.981445 Kb 0.062317 s |
| **Третий запрос** | 30432 Kb  0.019118 s | 26.375000 Kb 2.492477 s | 9.405273 Kb 45.764119 s |
| **Четвертый запрос** | 12624 kb  0.773233 s | 29.250000 Kb 0.015439 s | 43.804688 Kb 37.599624 s |
| **Пятый запрос** | 44400 kb  0.023516 s | 293.248047 Kb 0.007221 s | - |

В датасете MOOC ArangoDB показывает лучшую производительность по отношению к использованию памяти для большинства запросов. В то время как Dgraph отличается скоростью, ArangoDB использует гораздо меньше памяти при обработке этих типов запросов.

Scylla обладает достаточно высокой скоростью обработки запросов BFS и DFS, но она ниже, чем в ArangoDB и Dgraph.

Dgraph демонстрирует преимущество в производительности для второго и третьего запроса - скорость выполнения значительно выше, чем в других базах данных.

В таблице 8 представлены результаты для Elliptic++.

Таблица 8 – Результаты для Elliptic++ Transactions Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Запросы\База данных** | **Dgraph** | **ArangoDb** | **Scylla** |
| **BFS** | 9952 Kb | 285.098633 Kb 0.008803 s | 16.144531 Kb 0.169322 s |
| **DFS** | 0.039968 s | 285.862305 Kb 0.004760 s | 31.654297 Kb 0.067142 s |
| **Второй запрос** | 11264 Kb  0.004517 | 28.472656 Kb 0.213666 s | 4.981445 Kb 0.014819 s |
| **Третий запрос** | 14896 Kb  1.593489 s | 284.287109 Kb 4.781357 s | - |
| **Четвертый запрос** | 15264 Kb  1.317387 s | 286.571289 Kb 0.253495 s | - |
| **Пятый запрос** | 22416 Kb  0.024107 s | 577.100586 Kb 0.006219 s | - |

Исходя из результатов, можно сделать выводы: Dgraph демонстрирует превосходную скорость выполнения первого рекурсивного и второго запросов. Время обработки данных в Dgraph значительно ниже, чем в других базах данных.

ArangoDB отличается практически стабильным потреблением памяти при выполнении запросов, но скорость обработки ниже, за исключением 2-го и 5-го. Время выполнения второго запроса в ArangoDB уступает Dgraph и Scylla.

Scylla имеет более низкую скорость обработки запросов BFS и DFS, но второй запрос показал высокую эффективность.

Таблица 9 демонстрирует результаты выполнения запросов к набору данных California road network, позволяющие сравненить производительность баз данных Dgraph, ArangoDB и Scylla.

Таблица 9 – Результаты для California road network

|  |  |  |  |
| --- | --- | --- | --- |
| **Запросы\База данных** | **Dgraph** | **ArangoDb** | **Scylla** |
| **BFS** | 29680 Kb  0.053329 s | 26.312500 Kb  372.876247 s | 154.381836 Kb 19.340542 s |
| **DFS** |  | 287.267578 Kb  367.248468 s | 57.225586 Kb 1.190700 s |
| **Второй запрос** | - | - | - |
| **Третий запрос** | 33584 Kb  25.199382 s | 10.911133 Kb 32.426726 s | - |
| **Четвертый запрос** | - | - | - |
| **Пятый запрос** | 22560 Kb  0.020936 | 0.561646 Kb 0.088699 s | - |

В данном датасете Dgraph демонстрирует высокую производительность для запросов, связанных с графами, особенно первый рекурсивный запрос, но потребляет много памяти во всех запросах.

ArangoDB превосходит в использовании памяти при выполнении запросов BFS и DFS, но скорость выполнения ниже, чем у Dgraph.

Dgraph показывает преимущество в производительности "пятого запроса" - запрос обрабатывается быстрее, чем в ArangoDB.

Scylla отличается высокой скоростью обработки запросов BFS и DFS по сравнению с ArangoDB, но ее производительность для других типов запросов не была исследована.

Таблица 10 – Результаты для Stablecoin ERC20 Transactions Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Запросы\База данных** | **Dgraph** | **ArangoDb** | **Scylla** |
| **BFS** | 30912 Kb  0.016682 s | 32.653320 Kb 0.043762 s | 154.523438 Kb 0.021544 s |
| **DFS** |  | 32.708008 Kb 0.010981 s | 154.421875 Kb 0.015514 s |
| **Второй запрос** | 21856 Kb  0.021476 s | 0.010595 Kb 0.082903 s | 4.981445 Kb 0.235799 s |
| **Третий запрос** | 15840 Kb  4.833872 s | 67.112305 Kb 450.183318 s | - |
| **Четвертый запрос** | 37216 Kb  15.440171 s | 63.942383 Kb 3502.704748 s | - |
| **Пятый запрос** | 3584 Kb  0.002086 s | 32.762695 Kb 0.497944 s | - |

Получив результаты, можно сделать выводы:

Dgraph демонстрирует стабильно высокую скорость выполнения всех запросов. Время выполнения этих запросов в Dgraph значительно ниже, чем в других базах данных.

Scylla отличается высокой скоростью обработки запросов BFS и DFS. Время выполнения запросов BFS и DFS в Scylla ниже, чем в ArangoDB.

ArangoDB отличается практически стабильным потреблением памяти при выполнении запросов, но скорость обработки ниже.

Время выполнения запросов в ArangoDB значительно выше, чем в Dgraph, особенно для "третьего" и "четвертого" запроса.

**Заключение**

В результате практики поставленные задачи были выполнены. Все результаты запросов были получены, за исключением некоторых запросов в базе данных Scylla, для которых данные графовые запросы не подходят.

Можно сделать вывод, что для задач, где требуется высокая скорость обработки запросов, Dgraph является хорошим выбором, если ограничения по использованию памяти не критичны. ArangoDB может быть более подходящим вариантом, если требуется минимизировать потребление памяти, но скорость выполнения ниже чем у Dgraph. Scylla демонстрирует хорошую производительность для графовых запросов BFS и DFS, но ее функциональность для других типов запросов требует дополнительного исследования.

Теоретические знания и практические навыки, приобретенные в процессе практики, будут полезны в решении актуальных задачах в университете, работе. Результаты приведенной работы помогут при выборе базы данных для работы с набором данных с похожей структурой и запросами.

# СПИСОК ЛИТЕРАТУРЫ

1. Официальный сайт ЗАО «НОРСИ-ТРАНС». − URL: https://norsi-trans.ru/about-company/ (дата обращения: 27.07.2024).
2. MOOC User Action Dataset https://snap.stanford.edu/data/act-mooc.html (дата обращения: 30.07.2024).
3. Elliptic++ Transactions Dataset; [https://github.com/git](https://github.com/gitdisl/EllipticPlusPlus/tree/main/Transactions%20Dataset)[disl/EllipticPlusPlus/tree/main/Transactions%20Dataset](https://github.com/gitdisl/EllipticPlusPlus/tree/main/Transactions%20Dataset)(дата обращения: 30.07.2024).
4. California road network [https://snap.stanford.edu/d](https://snap.stanford.edu/data/roadNet-CA.html)[ata/roadNet-CA.html](https://snap.stanford.edu/data/roadNet-CA.html) (дата обращения: 30.07.2024).
5. Stablecoin ERC20 Transactions Dataset https://snap.stanford.edu/data/ERC20-stablecoins.html(дата обращения: 30.07.2024).