НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Курсовая работа

ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ СОВРЕМЕННЫХ АНАЛИТИЧЕСКИХ СУБД.

Узиков Александр Витальевич, БПМИ166

Научный руководитель:

Руководитель группы разработки СУБД ClickHouse, Яндекс

Миловидов Алексей Николаевич

Москва, 2019

АКТУАЛЬНОСТЬ

- Разработчикам СУБД важно знать сильные и слабые стороны конкурентов для более эффективной работы рад продуктом.
- Пользователям необходимо знать сильные и слабые стороны СУБД для того, чтобы выбрать наиболее подходящую.
- СУБД постоянно дорабатываются, а значит актуальные характеристики меняются.

ВСЕ ЛИ СУБД КОРРЕКТНО СРАВНИВАТЬ?

- Разные типы данных
- СУБД могут вести себя по разному в зависимости от числа узлов в кластере
- Разные сценарии работы

СЦЕНАРИЙ РАБОТЫ:OLTP

1. Обработка непрерывного потока запросов.
2. Эти запросы в основном простые: добавление или удаление строки, чте-
ние одной или нескольких строк, получение отдельного значения и т.д.
3. Основная задача - обработка большого числа запросов в секунду.
4. Высокие требования к консистентности данных.

СЦЕНАРИЙ РАБОТЫ:OLAP

1. Большая часть запросов на чтение, а не на запись.
2. Запросов мало, но они сложные и включают в себя обработку большого
массива данных.
3. Результат выполнения сильно меньшего размера, чем обработанные данные и помещается в оперативную память.
4. Консистентность важна меньше, чем при OLTP сценарии.
5. Короткие запросы могут обрабатываться сравнительно медленно.
6. Таблицы очень широкие.
7. Все данные хранятся в одной большой таблице.
8. При чтении вынимается сравнительно небольшое количество столбцов.

ИСПОЛЬЗУЕМЫЕ ДАННЫЕ

- Обфусцированные данные о поведении пользователей в сети
- Размер – 100 миллионов записей, каждая из которых имеет более 100 полей различных типов
-

ИСПОЛЬЗУЕМОЕ ОБОРУДОВАНИЕ

Ноутбук Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz, GeForce GTX 950M, 12Gb RAM
Облако Intel Core Processor (Haswell, no TSX), 66 Gb RAM

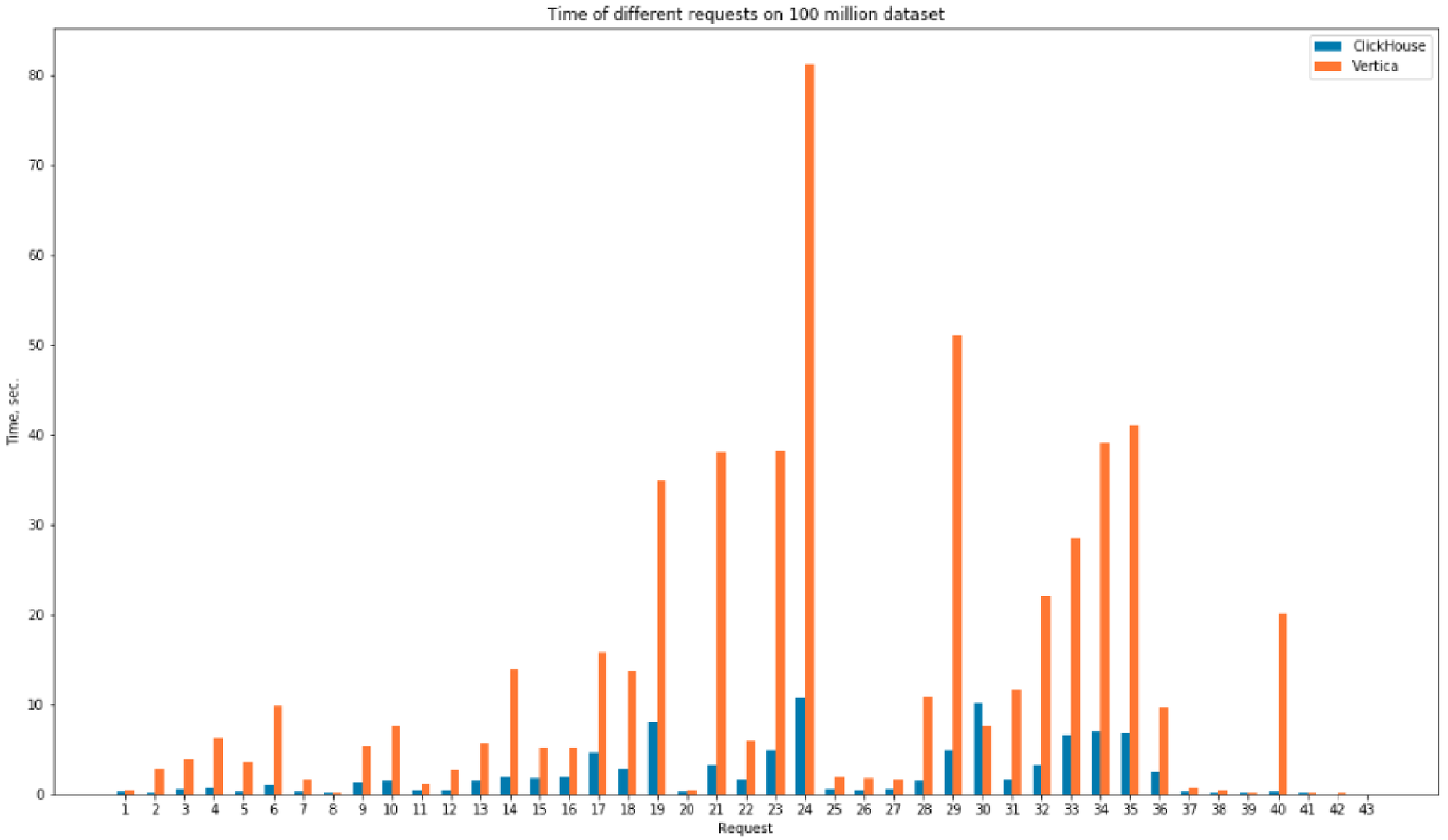
МЕТОД ИССЛЕДОВАНИЯ

Запуск на 43 аналитических запросах различной сложности, включающих в себя работу со строками, регулярными выражениями, группировкой по нескольким элементам и т. д.

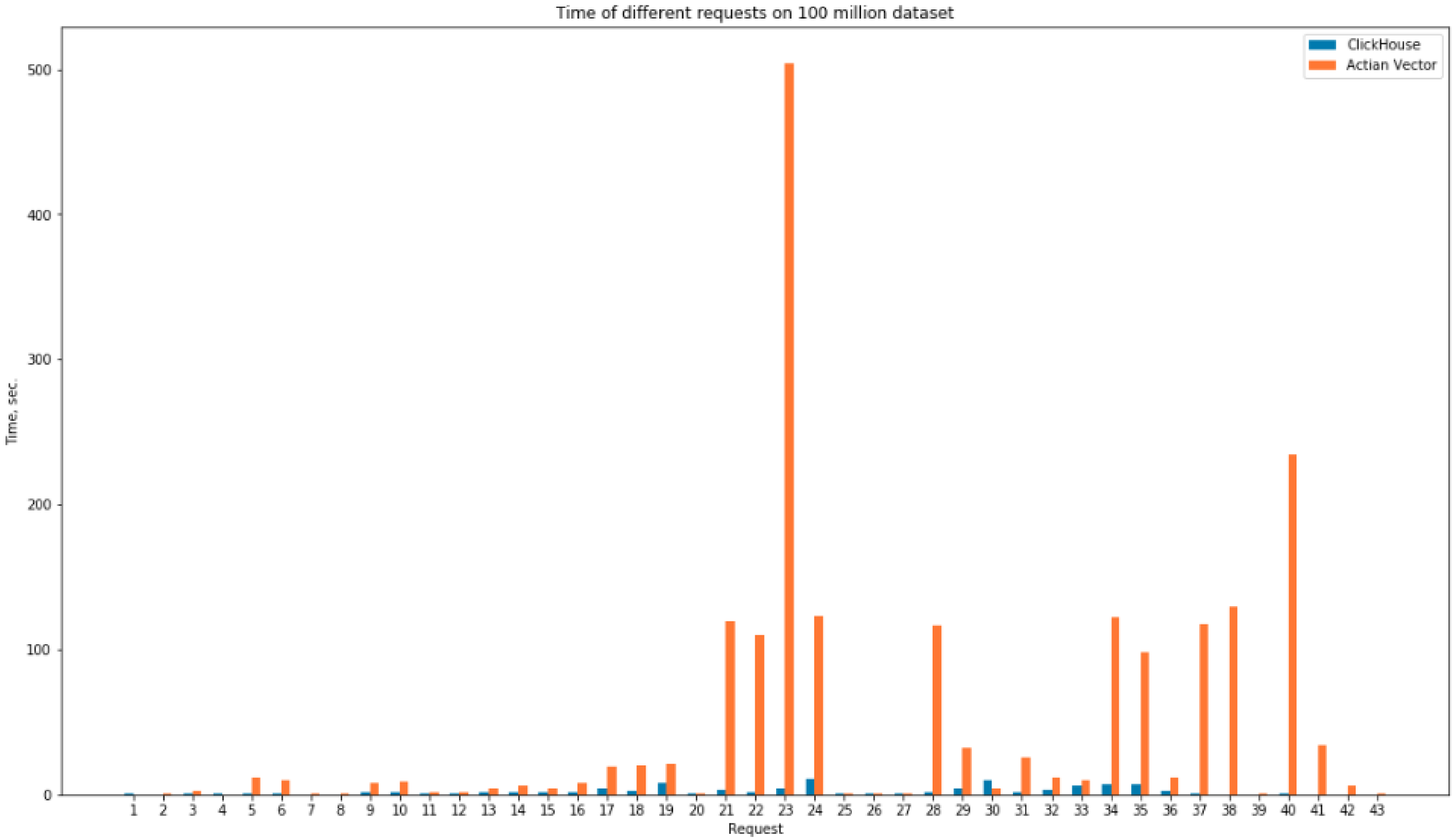
ПРЕДЫДУЩИЕ ИССЛЕДОВАНИЯ

Данное исследование - повторение исследование 2013 года и обновление его результатов.
Существует множество других бенчмарков, часть из которых коротко описана в курсовой работе.

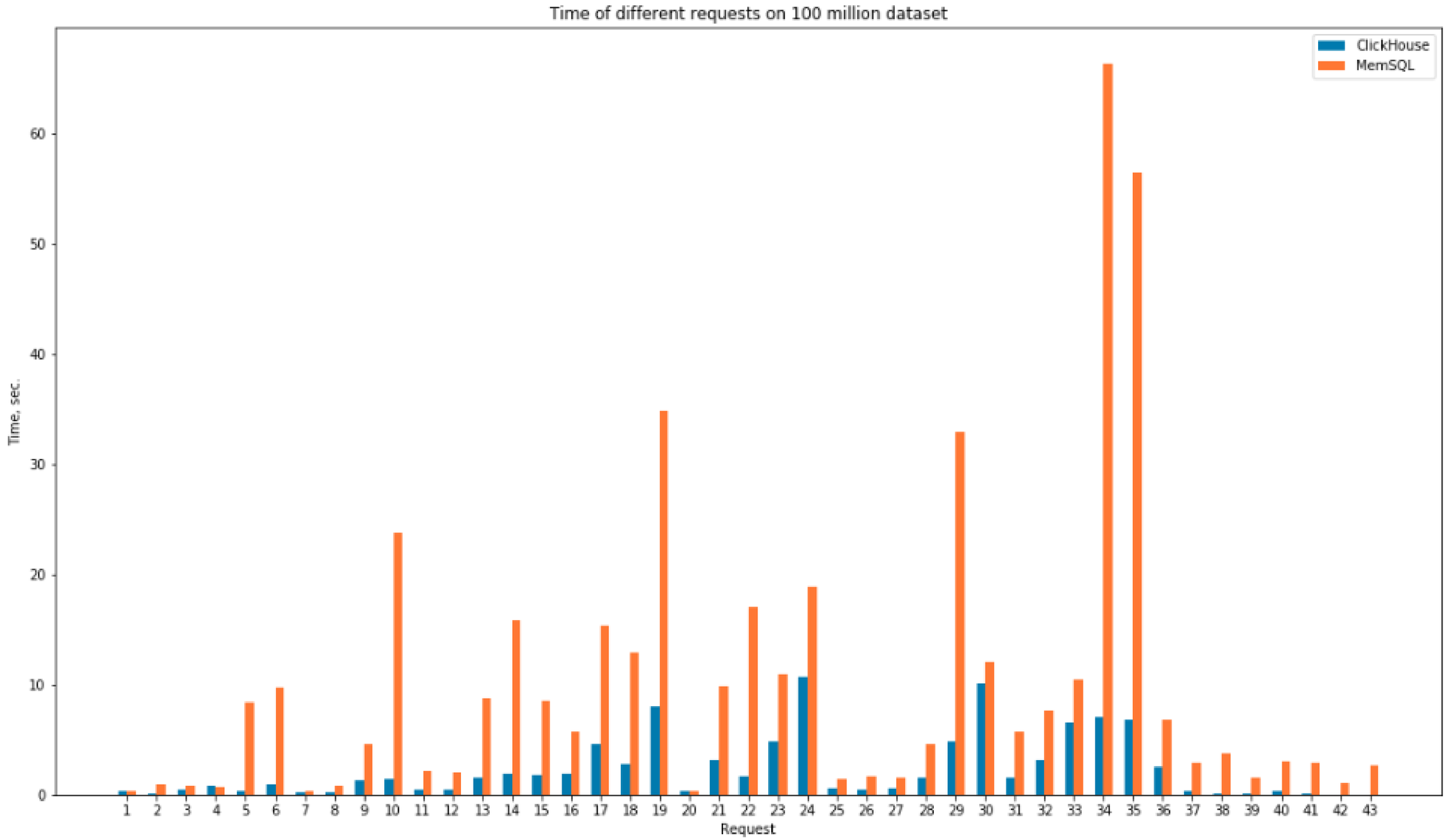
СРАВНЕНИЕ С VERTICA



СРАВНЕНИЕ С ACTIAN



СРАВНЕНИЕ С MEMSQL



РЕЗУЛЬТАТЫ

- ClickHouse показал наилучшие результаты почти на всех тестах
- В среднем ClickHouse работал в 2 раза быстрее Vertica, 7 раз. быстрее Actian Vector, в 4.5 раза быстрее MemSQL.
- Были выявлена слабая сторона Actian Vector: он медленно обрабатывает регулярные выражения.

СПИСОК ИСТОЧНИКОВ

- https://clickhouse.yandex/benchmark.html
- https://www.vertica.com/docs/9.2.x
- https://www.memsql.com/content/architecture/
- https://www.actian.com/wp-content
- http://www.tpc.org/tpch
- https://www.cs.umb.edu/ poneil/StarSchemaB.PDF
- https://amplab.cs.berkeley.edu/benchmark
- https://github.com/timescale/tsbs

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ