**Отчет по бенчмарку библиотек для работы с базами данных**

**Описание бенчмарка.** Целью бенчмарка было измерение производительности различных баз данных, используя следующие библиотеки: psycopg2 для работы с PostgreSQL, SQLite, DuckDB, Pandas, и SQLAlchemy. Бенчмарк включал выполнение четырех SQL-запросов, соответствующий код для каждого из них предоставлен ниже:

Запрос 1:

(SELECT cab\_type, count(\*) FROM trips GROUP BY 1;

Запрос 2:

SELECT passenger\_count, avg(total\_amount)

FROM trips

GROUP BY 1;

Запрос 3:

SELECT

passenger\_count,

extract(year from pickup\_datetime),

count(\*)

FROM trips

GROUP BY 1, 2;

Запрос 4:

SELECT

passenger\_count,

extract(year from pickup\_datetime),

round(trip\_distance),

count(\*)

FROM trips

GROUP BY 1, 2, 3

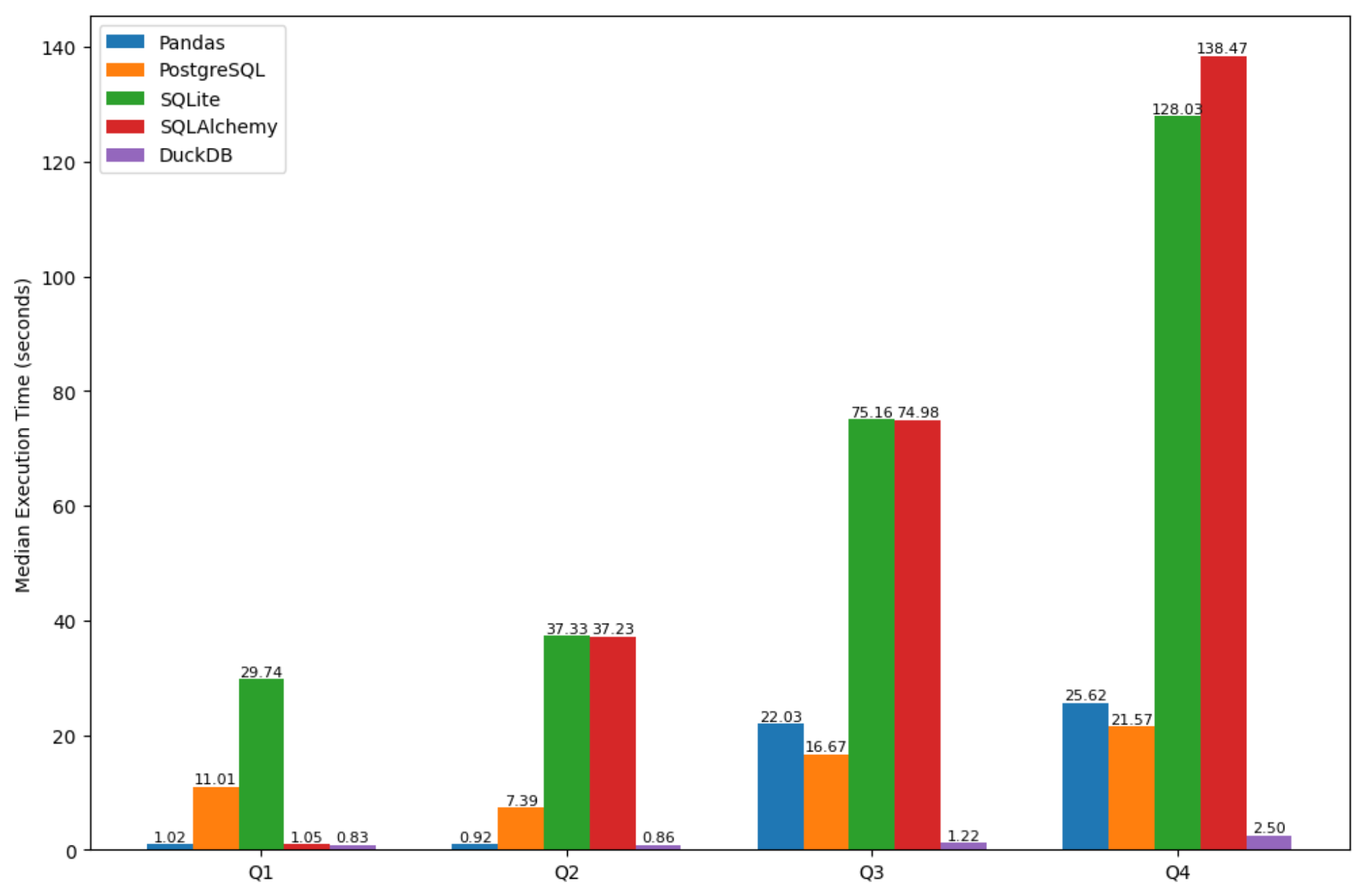
ORDER BY 2, 4 desc;

Результаты бенчмарка были выражены в медианных временах выполнения каждого запроса для каждой библиотеки.

**Датасет**. Для выполнения бенчмарка был использован набор данных по поездкам такси Yellow , который включает в себя информацию о времени и местах посадки и высадки, расстоянии поездки, детализации тарифов, типах тарифов, методах оплаты и отчетах водителей о количестве пассажиров. Данные были выбраны за 5 месяцев 2019 года. Общий объем данных составил 37 млн.строк

**Результаты и сравнительный анализ.**

Для каждой из библиотек было использовано 10 запусков на каждый запрос и измерено медианное время выполнения запросов. В итоге получили следующий график:



Из результатов видно, что наилучший результат показала библиотека DuckDB и Pandas. Далее по эффективности идет PostgreSQL. SQLite и SQLAlchemy показали значительно худшие результаты. Теперь проведем сравнительный обзор исследуемых библиотек.

* psycopg2 (PostgreSQL)

Библиотека psycopg2 для работы с PostgreSQL показала отличные результаты в скорости выполнения запросов. Медианные времена выполнения находятся в пределах от 7 до 21 секунды.

Из достоинств в использовании можно выделить простой и удобный API для взаимодействия с базой данных PostgreSQL. Однако, она требует более детальной работы с SQL запросами в сравнении с высокоуровневыми ORM-библиотеками.

* SQLite

SQLite также показала наиболее худшие результаты в скорости выполнения запросов. Медианные времена находятся в пределах от 29 до 128 секунд.

Из достоинств тут также можно выделить простоту и легкость в использовании. Также SQLite работает в рамках одного файла и не требует серверного взаимодействия, что обеспечивает некоторые преимущества в скорости для простых запросов.

* DuckDB

Скорость выполнения запросов: DuckDB показала впечатляющие результаты с медианными временами от 0.23 до 2.5 секунд. Данная библиотека также предоставляет простой SQL интерфейс, что делает ее удобной в использовании. Для объяснения полученного результата, можно предположить, что библиотека оптимизирована для аналитических запросов, что обеспечивает высокую производительность при агрегировании данных и выполнении сложных запросов.

* Pandas

Pandas также наряду с DuckDB показала отличную производительность для аналитических запросов, за исключением последнего запроса, где время выполнения существенно выше. Медианные времена находятся в пределах от 0.92 до 25.62 секунд. Однако время выполнения запросов в Pandas зависит от объема данных и сложности операций.

* SQLAlchemy

SQLAlchemy, как и SQLite показала один из низких время выполнения запросов. Медианные времена находятся в пределах от 28.76 до 138.47 секунд.

Из недостатков тут выделяется отсутствие оптимизации для конкретной базы что влияет на ее производительность.

**Выводы:**

SQLite и SQLAlchemy подходят для небольших проектов и прототипирования, где простота важнее высокой производительности.

Pandas хороша для работы с данными в памяти, но производительность может страдать при больших объемах данных.

DuckDB и psycopg 2 показали отличную производительность для аналитических запросов.