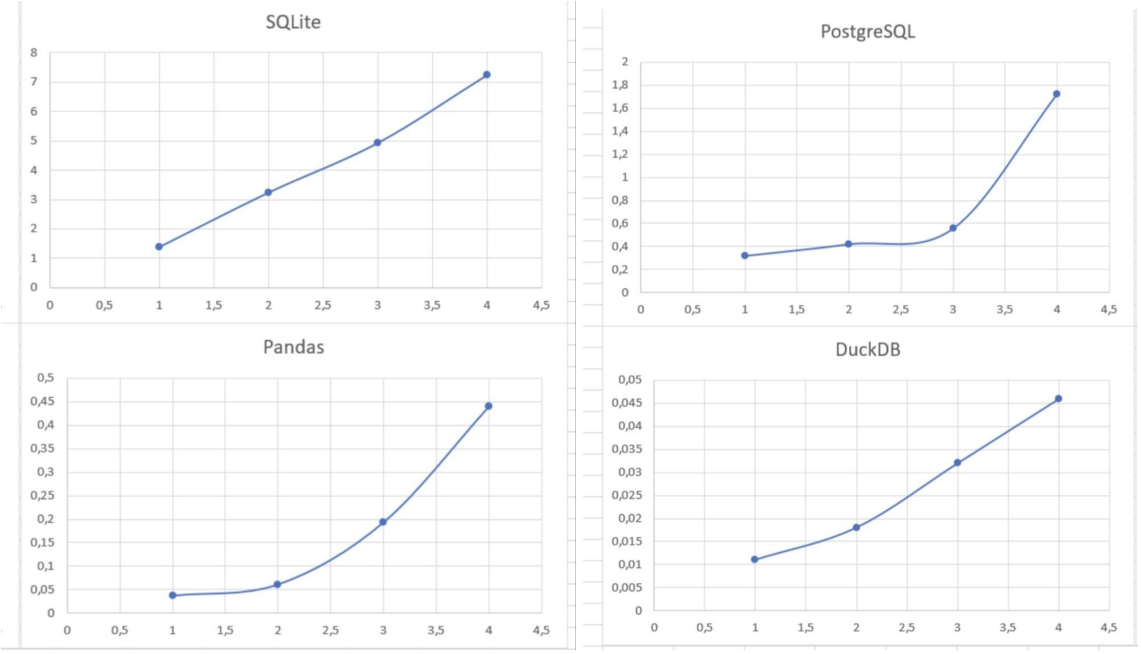
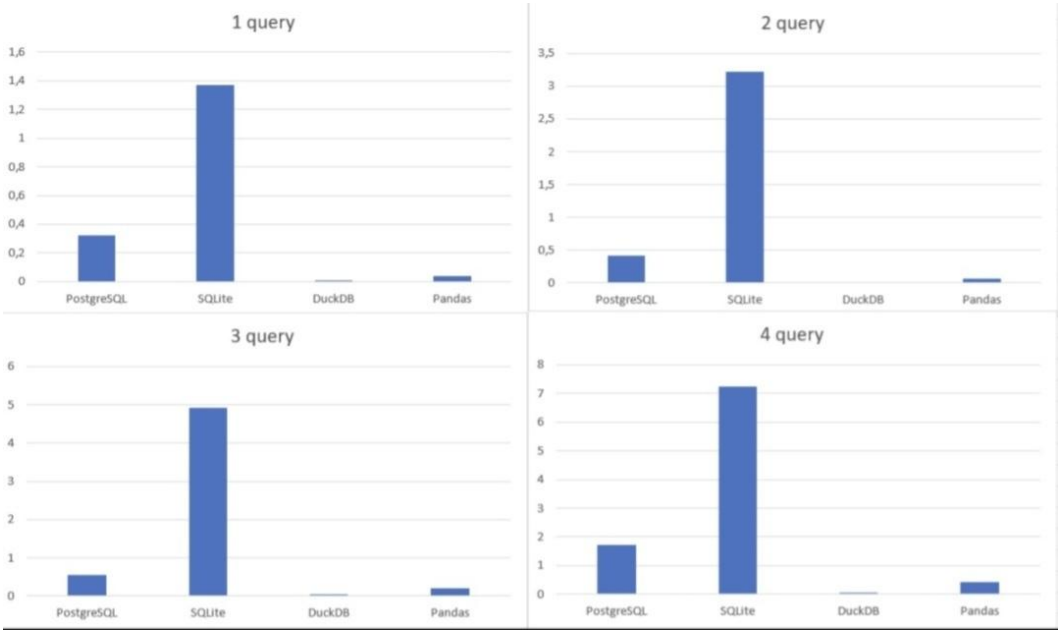
GitHub:
https://github.com/YoJoonghyuk/laba3.git





Postgres
PostgreSQL — это продвинутая РСУБД, которая является масштабируемым решением и позволяет обрабатывать большие объёмы данных. Поддерживает формат json. Скорость работы может падать во время проведения пакетных операций или выполнения запросов чтения.

Sqlite

SQLite — это библиотека, встраиваемая в приложение. Является автономной. Будучи файловой БД, она предоставляет отличный набор инструментов для более простой обработки любых видов данных в сравнении с серверными БД. Когда приложение использует SQLite, их связь производится с помощью функциональных и прямых вызовов файлов, содержащих данные (например, баз данных SQLite), а не какого-либо интерфейса, что повышает скорость и производительность операций. Так как база данных хранится в одном файле, это облегчает её перемещение. Несмотря на то, что SQLite является встроенной базой данных, при выполнении исчерпывающего анализа работает слишком медленно.

DuckDB
DuckDB - бессерверная система управления аналитической базой данных. Она достаточно быстра. Это достигается за счет векторизации выполнения запросов (ориентации на столбцы). Другие СУБД, упомянутые ранее, такие как SQLite, PostgreSQL обрабатывают каждую строку последовательно. Именно за счет этого по производительности DuckDB выигрывает. Также DuckDB проста в использовании. Более того, DuckDB не имеет внешних зависимостей и серверного программного обеспечения, которое нужно устанавливать, обновлять и поддерживать. DuckDB — полностью встроенная система, что обеспечивает дополнительное преимущество — высокоскоростную передачу данных в базу данных и из нее. Она поддерживает сложные запросы в SQL. Но может иметь некоторые ограничения в поддержке SQL-функций.

Pandas
Pandas — одна из самых популярных библиотек Python. DataFrame интуитивно понятен и оснащен продвинутыми API для выполнения задач по работе с данными. Однако библиотека Pandas работает медленно на больших наборах данных. Основная причина заключается в том, что библиотека не была создана для работы на нескольких ядрах. Pandas использует только одно ядро процессора за раз для выполнения задач по манипулированию данными и хранит данные в памяти целиком.

Вывод :
DuckDB выигрывает по скорости выполнения запросов, благодаря хранилищу на основе столбцов и векторизованному выполнению. В целом pandas обгоняет Postgres по времени выполнения запросов. Выборка столбцов в pandas происходит эффективно с O(1) времени, поскольку фрейм данных уже сохранен в памяти. Pandas является мощным инструментом для изученных задач анализа данных. Однако pandas имеет свои ограничения. В pandas данные хранятся в памяти, и будет сложно загрузить CSV-файл, занимающий более половины памяти системы. Pandas был создан для манипулирования данными, и его сильная сторона заключается в сложных операциях по анализу. А Postgres и другие языки, основанные на SQL, были созданы для управления базами данных и предоставления пользователям удобного способа доступа к данным и их извлечения. В свою очередь SQLite демонстрировала стабильно самую низкую производительность среди других 3 библиотек. Это может быть связано с особенностями алгоритмов хранения и индексации. Однако SQLite также имеет свои преимущества в виде простоты, легковесности и отсутствия необходимости установки и конфигурации сервера баз данных, работающего локально.