Документирование выполненных действий

Предоставлено командой «Дети-GPT»

**1. Разработанное приложение**

**1.1 Описание приложения**

Разработано Python-приложение для анализа логов веб-серверов.

Приложение включает:

* Модуль загрузки данных из лог-файлов в PostgreSQL.
* Модуль визуализации статистики (графики запросов, ошибок, таблицы IP-адресов).
* Модуль экспорта данных в CSV и Parquet.
* Графический интерфейс на PySide6 для удобного управления процессами.

**1.2 Требования для запуска**

* Python 3.12+.
* Установленные модули:
* pip install pandas psycopg2-binary PySide6 matplotlib chardet openpyxl
* Настроенная СУБД PostgreSQL с параметрами из CONFIG в коде (хост, порт, пароль и т.д.).

**1.3 Возможности приложения**

* Загрузка одиночных файлов или папки с логами.
* Автоматическая проверка уникальности записей через хеширование.
* Визуализация:
  + Графики запросов/ошибок по дням.
  + Таблица IP-адресов с количеством обращений по дням.
* Экспорт в CSV/Parquet для дальнейшего анализа.

**2. Интерфейс**

**2.1 Графический интерфейс**

* Реализован на PySide6 (аналог Qt для Python).
* Основные элементы:
  + Кнопки для загрузки файлов/папок.
  + Прогресс-бар и статус-бар.
  + Таблица для отображения данных IP-адресов.
  + Кнопки экспорта и визуализации.

**2.2 Пример использования**

* Запустить программу.
* Нажать "Загрузить папку с логами" и выбрать директорию.
* После обработки использовать кнопки для построения графиков или экспорта.

**3. Структура базы данных**

**3.1 СУБД**

Выбрана PostgreSQL (реляционная БД) для:

* Эффективного хранения структурированных логов.
* Поддержки сложных запросов и агрегаций.

**3.2 Схема БД**

Модель "звезда":

* Таблица ip\_addresses:
  + ip\_address (PK), total\_requests, is\_suspicious.
* Таблица requests:
  + Связь с ip\_addresses через ip\_id, поля: datetime, method, status\_code, is\_error.
* Материализованное представление daily\_stats:
  + Агрегированная статистика по дням (запросы, ошибки, время ответа).

**3.3 Настройка**

Параметры подключения в CONFIG (хост, порт, имя БД).

Скрипты для создания схемы (create\_schema) и очистки (clear\_database).

**4. Производительность**

* Обработка 1 лога и запись его в базу данных занимает 2.5 минуты.
* Запись 20 логов и его запись в БД составляет около часа.
* Экспорт в CSV и Parquet занимает около 10 секунд.

**5. Технологии и алгоритмы**

* Язык: Python 3.12.
* Библиотеки:
  + pandas для обработки данных.
  + psycopg2 для работы с PostgreSQL.
  + matplotlib для визуализации.
* Алгоритмы:
  + Хеширование для проверки уникальности логов.
  + Регулярные выражения для парсинга строк логов.

**6. Особенности реализации**

* Разработан оптимизированный анализатор логов веб-сервера с графическим интерфейсом (GUI) на PySide6.
* Реализована быстрая обработка больших объемов данных за счет:
* Многопоточной загрузки и парсинга логов.
* Хеширования записей для исключения дубликатов.
* Пакетной вставки в PostgreSQL (до 20 млн записей за раз).
* Визуализация данных: графики запросов/ошибок, таблицы с IP-статистикой.
* Экспорт в CSV и Parquet для дальнейшего анализа.

**7. Формат выгружаемых данных**

* CSV:
  + Столбцы: IP, дата, метод, статус, ошибки и т.д.
* Parquet:
  + Оптимизирован для анализа в датафреймах (поддержка типов данных).