**ДИПЛОМНЫЙ ПРОЕКТ**

Курс: «Python-разработчик»

Тема: «**Анализ и сравнение различных способов обработки и хранения больших данных: Pandas и Dask**»

Выполнил: Магомедов Б.К.

**Оглавление**

[**1. ВВЕДЕНИЕ 2**](#_Toc182391608)

[**2. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ 3**](#_Toc182391609)

[**3. МЕТОДЫ И ПОДХОДЫ К РАЗРАБОТКЕ 5**](#_Toc182391610)

[**4. ОБЗОР ПОПУЛЯРНЫХ ИНСТРУМЕНТОВ ДЛЯ РАЗРАБОТКИ ПРИЛОЖЕНИЙ НА PYTHON ДЛЯ РАБОТЫ С БОЛЬШИМИ ДАННЫМИ. 5**](#_Toc182391611)

[**5. ПРОЕКТИРОВАНИЕ ПРИЛОЖЕНИЯ ДЛЯ ОБРАБОТКИ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ PANDAS И DASK. 7**](#_Toc182391612)

[**6. РАЗРАБОТКА В СООТВЕТСТВИИ С СОЗДАННОЙ ДОКУМЕНТАЦИЕЙ 8**](#_Toc182391613)

[**7. АНАЛИЗ И ИНТЕРПЕТАЦИЯ РЕЗУЛЬТАТОВ 8**](#_Toc182391614)

[**8. ЗАКЛЮЧЕНИЕ 9**](#_Toc182391615)

1. **ВВЕДЕНИЕ**

**Обоснование выбора темы**

1. Актуальность анализа больших данных: В современном мире, с учетом постоянного роста объемов данных, их эффективная обработка и хранение стали одними из важнейших задач. Обработка больших данных находит применение в различных областях, таких как анализ рыночных тенденций, предсказание пользовательского поведения, научные исследования и др. Важно понимать, какие инструменты обработки данных лучше справляются с задачами на больших объемах данных.
2. Растущая потребность в сравнении инструментов: Существует множество библиотек и платформ для обработки больших данных, среди которых особенно выделяются Pandas и Dask. Каждая из этих технологий обладает уникальными преимуществами и ограничениями. Для разработчиков и аналитиков важно понимать, в каких случаях целесообразно использование каждого из этих инструментов, чтобы выбрать оптимальное решение для конкретных задач.
3. Практическая значимость исследования: Работа над созданием и сравнением приложений на базе Pandas и Dask позволит получить практические навыки и опыт в работе с большими данными и их обработкой. Это позволит выявить сильные и слабые стороны каждой технологии, что является ценным для бизнеса и разработки более эффективных продуктов.
4. Личный интерес и перспективы: Изучение и освоение различных подходов к обработке и хранению больших данных предоставляет возможность совершенствоваться в области аналитики данных и разработки, а также открывает перспективы для карьерного роста. Этот опыт востребован на рынке, где специалисты по большим данным пользуются высоким спросом.

Таким образом, тема дипломной работы «Анализ и сравнение различных способов обработки и хранения больших данных: Pandas и Dask» представляется актуальной и значимой, учитывая ее востребованность в профессиональной среде, а также важность получения практических навыков в этой области.

**Определение цели и задач исследования**

Цель исследования: Разработать приложения для обработки больших данных с использованием Pandas и Dask и провести сравнительный анализ их производительности и эффективности.

Задачи исследования:

* Провести обзор фреймворков Pandas и Dask, выделить особенности их работы и области применения.
* Разработать сценарии и критерии для тестирования производительности каждого инструмента на различных объемах данных.
* Создать приложения для обработки данных на основе каждого инструмента и протестировать их на выборках данных.
* Провести анализ полученных данных, выделить сильные и слабые стороны каждого инструмента, сделать выводы и рекомендации по их применению.

### ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

#### Обзор основных понятий в области обработки и хранения больших данных с использованием Pandas и Dask:

1. **Большие данные (Big Data):** Это массивы данных, которые слишком объемны, чтобы быть обработанными традиционными методами и требуют специализированных технологий и подходов. Характеристики больших данных включают большой объем, разнообразие форматов, высокую скорость поступления и рост данных.
2. **Фреймворк (Framework):** Программная платформа, предоставляющая готовые компоненты и инструменты для обработки данных. В контексте анализа больших данных основными фреймворками являются Pandas и Dask, которые позволяют реализовывать эффективную обработку и анализ данных.
3. **Pandas:** Библиотека Python, предоставляющая средства для работы с данными в табличном формате. Pandas широко используется для обработки небольших и средних объемов данных, когда можно работать в рамках одного компьютера, так как она использует только одно ядро процессора.
4. **Dask:** Библиотека Python для параллельных вычислений, которая позволяет обрабатывать данные, превышающие объем оперативной памяти. Dask масштабируется от одного компьютера до распределенной среды, и его API совместим с Pandas, что делает его удобным для пользователей, знакомых с Pandas, но нуждающихся в работе с большими объемами данных.
5. **Параллелизм (Parallelism):** Метод, позволяющий выполнять несколько задач одновременно, что увеличивает производительность при обработке данных. Pandas ограничен параллелизмом на одном ядре, тогда как Dask могут работать на множестве ядер и узлов.
6. **Эффективность обработки данных (Data Processing Efficiency):** Оценка производительности инструмента или платформы при обработке данных, учитывающая такие метрики, как время выполнения, потребление памяти и масштабируемость. Эффективность обработки данных является ключевым фактором при выборе платформы для анализа больших данных.
7. **Скалируемость (Scalability):** Способность системы увеличивать производительность при увеличении объема данных или числа вычислительных ресурсов. Dask обеспечивают высокую скалируемость за счет распределенной архитектуры, в отличие от Pandas, который ориентирован на одномашинные расчеты.
8. **Временная сложность (Time Complexity):** Мера, описывающая, как количество операций по обработке данных увеличивается с ростом объема данных. Важно учитывать временную сложность алгоритмов, используемых в Pandas и Dask, так как это влияет на выбор оптимального инструмента для конкретной задачи.
9. **Время отклика (Response Time):** Время, необходимое инструменту для выполнения операций над данными. Эта метрика важна для оценки производительности и пригодности инструмента для задач, требующих обработки в реальном времени или близкой к нему.
10. **API совместимость (API Compatibility):** Способность библиотеки или фреймворка взаимодействовать с другими инструментами, сохраняя аналогичный интерфейс. Dask, например, использует API, совместимый с Pandas, что упрощает переход для пользователей.
11. **МЕТОДЫ И ПОДХОДЫ К РАЗРАБОТКЕ**

**Выбор и интеграция инструментов для работы с большими данными**

1. Выбор инструментов: Для реализации дипломного проекта, который направлен на обработку и анализ больших данных, необходимо рассмотреть ключевые фреймворки — Pandas и Dask. Каждый из них имеет свои особенности, которые делают его более подходящим для определенных задач:
   * Pandas подходит для обработки небольших объемов данных и анализа в рамках одного компьютера.
   * Dask полезен при обработке данных, которые могут превышать объем оперативной памяти, но при этом поддерживает знакомый пользователям Pandas API.
2. Интеграция инструментов: Для анализа и сравнения каждого инструмента будет разработан скрипт обработки данных, используя соответствующие API. Включение каждой технологии будет состоять из:
   * Загрузки данных и их первичной очистки.
   * Реализации одинаковых операций над данными (таких как фильтрация, агрегация, и преобразование), чтобы получить сопоставимые результаты.

**Обеспечение безопасности и стабильности системы**

1. Аутентификация и авторизация: Для ограничения доступа к результатам и данным пользователей можно добавить систему аутентификации. Это гарантирует, что только авторизованные пользователи могут отправлять данные и просматривать результаты их обработки.
2. Обработка ошибок и логирование: Для повышения стабильности проекта важно реализовать обработку ошибок, особенно при работе с большими объемами данных, что может вызывать исключения. Логирование операций поможет отслеживать и анализировать ошибки, а также проводить мониторинг производительности.

**4. ОБЗОР ПОПУЛЯРНЫХ ИНСТРУМЕНТОВ ДЛЯ РАЗРАБОТКИ ПРИЛОЖЕНИЙ НА PYTHON ДЛЯ РАБОТЫ С БОЛЬШИМИ ДАННЫМИ.**

Pandas — это библиотека на языке Python, которая предоставляет высокоуровневые структуры данных и инструменты для анализа данных. Она позволяет работать с данными в табличном виде и поддерживает широкий спектр операций, таких как фильтрация, агрегация, и манипуляция данными, обеспечивая простоту и удобство для аналитиков и разработчиков.

**Основные возможности:**

* DataFrame и Series: Основные структуры данных, предоставляющие табличный формат (DataFrame) и одномерный массив (Series) для работы с индексированными данными.
* Гибкие операции с данными: Поддержка большого набора встроенных функций для обработки данных, включая фильтрацию, агрегацию, объединение таблиц (join, merge), работу с пропусками в данных и другие манипуляции.
* Интеграция с другими библиотеками: Поддержка интеграции с такими инструментами, как NumPy, Matplotlib, Scikit-Learn, что делает Pandas удобным для работы в рамках анализа данных и машинного обучения.
* Чтение и запись данных: Pandas поддерживает широкий набор форматов файлов, включая CSV, Excel, SQL базы данных, JSON, что облегчает импорт и экспорт данных.

**Особенности:**

* Идеален для локальных операций: Pandas подходит для работы с данными, которые можно полностью загрузить в память компьютера, что делает его удобным для анализа небольших и средних объемов данных.
* Высокая производительность для небольших данных: Благодаря использованию оптимизированных структур данных и поддержке интеграции с библиотекой NumPy, Pandas обеспечивает высокую производительность для обработки небольших наборов данных.
* Простой и интуитивно понятный API: Интерфейс Pandas интуитивен и легко освоим, что делает его популярным выбором среди аналитиков данных и разработчиков, особенно тех, кто работает в Python.

Dask: Платформа, ориентированная на масштабируемые параллельные вычисления, которая поддерживает работу с большими наборами данных за счет выполнения операций как на локальном уровне, так и на уровне кластера.

* + Основные возможности:
    - API, совместимый с Pandas, что делает Dask удобным для пользователей, уже знакомых с Pandas.
    - Возможность работы с данными, превышающими объем оперативной памяти.
    - Поддержка распределенных вычислений и горизонтального масштабирования.
  + Особенности: Идеален для задач, которые не помещаются в память, но могут быть решены с использованием параллельных операций.

**5. ПРОЕКТИРОВАНИЕ ПРИЛОЖЕНИЯ ДЛЯ ОБРАБОТКИ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ PANDAS И DASK.**

**Планирование и анализ требований**

* Цель: Разработать приложение, способное обрабатывать и хранить большие объемы данных, используя различные инструменты: Pandas и Dask. Сравнить эффективность этих подходов при выполнении задач анализа данных.
* Выбор инструментов: Для обработки данных выбраны Pandas и Dask, что позволит оценить возможности и ограничения каждого инструмента.
* Определение структуры приложения: Архитектура включает интерфейс для загрузки и управления данными, бэкенд для обработки данных с использованием выбранного фреймворка, а также хранилище для хранения данных, метаданных и результатов анализа.

**Разработка прототипа**

* Реализация базового функционала: Создается минимальная версия приложения с простейшими функциями для загрузки данных, обработки с использованием Pandas и Dask, а также для отображения результатов. Этот прототип позволит выявить подходящие для проекта решения и проверить функциональность основных компонентов.

**Разработка ключевого функционала**

1. *Pandas*: Применяется для обработки данных небольшого объема или при работе с данными, которые полностью умещаются в оперативной памяти.
2. *Dask*: Настраивается для работы с данными, которые не умещаются в оперативной памяти, для оптимизации и параллельной обработки данных.

**Основные требования**

* Загрузка и обработка данных: Пользователь должен иметь возможность загружать наборы данных для анализа. Приложение должно поддерживать обработку данных с использованием Pandas и Dask, что позволит сравнить время обработки.

**Технические требования**

* Бэкенд: Python с использованием Flask или FastAPI для создания API и интеграции с фреймворками Pandas и Dask.
* Интеграция с фреймворками для обработки данных: Pandas и Dask для реализации и тестирования различных методов обработки данных.

**6.** **РАЗРАБОТКА В СООТВЕТСТВИИ С СОЗДАННОЙ ДОКУМЕНТАЦИЕЙ**

**Разработка**

* Настроены маршруты для обработки загрузки данных, выполнения задач обработки с использованием Pandas и Dask.

**Интеграция фреймворков для обработки данных**:

* Pandas и Daskбыли подключены и настроены для выполнения различных задач обработки данных. В зависимости от объема данных и требований, приложение выбирает наиболее подходящий инструмент.
* Реализован функционал для сравнения производительности Pandas и Dask. Приложение отображает метрики времени обработки, что позволяет пользователю оценить их эффективность.

**7. АНАЛИЗ И ИНТЕРПЕТАЦИЯ РЕЗУЛЬТАТОВ**

**Сравнение фреймворков**

В ходе разработки были протестированы фреймворки для обработки и хранения больших данных, такие как Pandas и Dask, на одном и том же наборе данных, чтобы оценить их производительность и удобство использования.

Основные метрики, использованные для оценки производительности каждого фреймворка, включали:

* Время обработки: время, необходимое для завершения анализа данных.

**Интерпретация результатов**

* Pandas:
  + Показал отличные результаты при обработке небольших и средних объемов данных.
  + Время обработки было минимальным, но при увеличении объема данных наблюдалось резкое увеличение использования памяти, что ограничивало его масштабируемость.
  + Рекомендуется для проектов с небольшими наборами данных, где требуется быстрая обработка и простой синтаксис.
* Dask:
  + Обеспечил хорошую производительность при работе с большими наборами данных.
  + Время обработки увеличивалось, но было значительно лучше, чем у Pandas при увеличении нагрузки.
  + Рекомендуется для проектов, где необходимо работать с объемами данных, превышающими возможности памяти на одном узле.

**Рекомендации по выбору фреймворка в зависимости от проекта**

* Для проектов с небольшими и средними объемами данных, где требуется быстрая разработка и простота в использовании, стоит выбирать Pandas.
* Если ожидается работа с большими объемами данных, то лучше использовать Dask.

**8. ЗАКЛЮЧЕНИЕ**

**Обзор выполненной работы**

В ходе проекта была успешно завершена разработка приложения для обработки и анализа больших данных с использованием различных фреймворков: Pandas и Dask. Приложение включает функционал загрузки данных, их обработки с использованием выбранного фреймворка и отображения результатов анализа.

Каждый из фреймворков был протестирован на одном наборе данных, и проведенный анализ продемонстрировал их сильные и слабые стороны. Pandas продемонстрировал отличную производительность для небольших наборов данных, но с ограничениями по масштабируемости. Dask показал хорошие результаты при работе с более крупными объемами данных, обеспечивая оптимальное использование памяти.

**Дальнейшие планы**

В качестве дальнейших шагов планируется создание сводной статистики по всем использованным фреймворкам, с учетом всех метрик эффективности, что позволит пользователям принимать более обоснованные решения о выборе инструмента для обработки данных в зависимости от специфики их проекта.

Также планируется добавление новых функций в приложение, таких как:

* Сохранение истории загрузок и результатов анализа, что упростит пользователям отслеживание прогресса.
* Расширение функциональности приложения за счет добавления новых методов обработки данных и интеграции с другими инструментами для анализа и визуализации данных.

Таким образом, проект продемонстрировал успешное применение современных фреймворков для обработки больших данных.