

MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică și Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susținere
Șef departament:
FIODOROV Ion dr., conf.univ.

„___” _____ 2024

DEZVOLTAREA UNUI SISTEM AUTOMATIZAT
PENTRU ANALIZA DATELOR DE AFACERI
Proiect de master

Student: _____ **Farid Salimzade, TIA-221M**
Coordonator: _____ **Cunev Veaceslav, lect. univ.**
Consultant: _____ **Cojocaru Svetlana, asist.univ.**

Chișinău, 2024

ADNOTARE

Acest proiect de teză este dedicat creării unui sistem automat pentru analiza datelor de afaceri folosind limbajul de programare Python. În lumea modernă, volumul datelor de afaceri este în continuă creștere, ceea ce necesită instrumente eficiente pentru analizarea acestora și luarea deciziilor. Scopul acestui proiect este dezvoltarea unui sistem care să permită companiilor să automatizeze procesul de analiză a datelor, să reducă costurile de timp și să îmbunătățească calitatea deciziilor luate.

Proiectul a inclus o trecere în revistă a principalelor metode de analiză a datelor în afaceri și a soluțiilor existente pe piață. Pe baza acestei analize, au fost selectate cele mai potrivite instrumente și tehnologii pentru dezvoltarea sistemului. Biblioteci și instrumente Python specializate, cum ar fi Pandas, NumPy și Apache Spark, au fost folosite pentru a procesa cantități mari de date. Pentru a vizualiza rezultatele analizei, au fost alese bibliotecile Matplotlib și Seaborn.

Pentru a stabili problema și a proiecta sistemul, au fost determinate cerințele pentru funcționalitatea sistemului și arhitectura acestuia. Au fost identificate principalele etape ale procesării datelor și metodelor de vizualizare care țin cont de specificul analizei de afaceri. Pentru a asigura fiabilitatea și scalabilitatea sistemului, au fost dezvoltați algoritmi și arhitectură adecvate.

Partea finală a lucrării a implicat implementarea sistemului, inclusiv crearea de scripturi și aplicații pentru colectarea, prelucrarea și analiza datelor. Au fost dezvoltați algoritmi pentru lucrul cu volume mari de date și a fost implementată funcționalitatea de vizualizare a rezultatelor analizei. Sistemul a fost apoi testat și evaluat în funcție de criteriile de performanță și conformitate.

La încheierea tezei au fost rezumate rezultatele studiului, s-au tras concluzii cu privire la dezvoltarea și implementarea sistemului și s-au dat recomandări pentru dezvoltarea și utilizarea ulterioară a acestuia. Teza include o listă de surse utilizate și aplicații cu codul sursă al scripturilor și aplicațiilor dezvoltate, exemple de date și rezultate ale analizelor, precum și materiale suplimentare care confirmă funcționalitatea și eficacitatea sistemului.

ABSTRACT

This thesis project is devoted to the creation of an automated system for analyzing business data using the Python programming language. In the modern world, the volume of business data is constantly growing, which requires effective tools for analyzing it and making decisions. The goal of this project is to develop a system that will allow companies to automate the process of data analysis, reduce time costs and improve the quality of decisions made.

The project included a review of the main methods of data analysis in business and existing solutions on the market. Based on this review, the most suitable tools and technologies for system development were selected. Specialized Python libraries and tools such as pandas, NumPy and Apache Spark were used to process large amounts of data. To visualize the analysis results, the Matplotlib and Seaborn libraries were chosen.

To set the problem and design the system, requirements for the functionality of the system and its architecture were determined. The main stages of data processing and visualization methods that take into account the specifics of business analysis were identified. To ensure the reliability and scalability of the system, appropriate algorithms and architecture have been developed.

The final part of the work involved the implementation of the system, including the creation of scripts and applications for collecting, processing and analyzing data. Algorithms for working with large volumes of data were developed, and functionality for visualizing analysis results was implemented. The system was then tested and evaluated against performance and compliance criteria.

At the conclusion of the thesis, the results of the study were summed up, conclusions were drawn about the development and implementation of the system, and recommendations were given for its further development and use. The thesis includes a list of sources used and applications with the source code of the developed scripts and applications, examples of data and analysis results, as well as additional materials confirming the functionality and effectiveness of the system.

АННОТАЦИЯ

Данный дипломный проект посвящен созданию автоматизированной системы для анализа данных бизнеса с использованием языка программирования Python. В современном мире объемы бизнес-данных постоянно растут, что требует эффективных инструментов для их анализа и принятия решений. Целью данного проекта является разработка такой системы, которая позволит компаниям автоматизировать процесс анализа данных, сократить временные затраты и повысить качество принимаемых решений.

В рамках проекта был проведен обзор основных методов анализа данных в бизнесе и существующих решений на рынке. На основе этого обзора были выбраны наиболее подходящие инструменты и технологии для разработки системы. Для обработки больших объемов данных были использованы специализированные библиотеки и инструменты Python, такие как pandas, NumPy и Apache Spark. Для визуализации результатов анализа были выбраны библиотеки Matplotlib и Seaborn.

Для постановки задачи и проектирования системы были определены требования к функционалу системы и ее архитектуре. Были выделены основные этапы обработки данных и методы визуализации, учитывающие специфику бизнес-анализа. Для обеспечения надежности и масштабируемости системы были разработаны соответствующие алгоритмы и архитектура.

В финальной части работы была проведена реализация системы, включая создание скриптов и приложений для сбора, обработки и анализа данных. Были разработаны алгоритмы для работы с большими объемами данных, а также реализован функционал визуализации результатов анализа. Затем система была протестирована и оценена по критериям производительности и соответствия требованиям.

В заключении дипломной работы были подведены итоги исследования, сделаны выводы о разработке и реализации системы, а также даны рекомендации по ее дальнейшему развитию и использованию. Дипломная работа включает в себя список использованных источников и приложения с исходным кодом разработанных скриптов и приложений, примерами данных и результатов анализа, а также дополнительными материалами, подтверждающими работоспособность и эффективность системы.

СОДЕРЖАНИЕ

1. Введение.....	9
1.1 Обоснование выбора темы.....	11
1.2 Цели и задачи исследования.....	12
1.3 Обзор основных методов анализа данных в бизнесе.....	13
2. Обзор существующих решений.....	14
2.1 Анализ существующих инструментов и платформ для анализа данных.....	16
2.2 Оценка их преимуществ и недостатков.....	17
2.3 Выбор основных инструментов и технологий для разработки.....	17
3. Постановка задачи и проектирование системы.....	20
3.1 Описание требований к системе.....	23
3.2 Проектирование архитектуры системы.....	26
3.3 Определение основных этапов обработки данных и методов визуализации.....	28
4. Реализация системы.....	30
4.1 Создание скриптов и приложений для сбора, обработки и анализа данных.....	32
4.2 Разработка алгоритмов для работы с большими объемами данных.....	34
4.3 Реализация функционала визуализации результатов анализа.....	34
5. Тестирование и оценка качества системы.....	36
5.1 Проведение тестирования разработанных компонентов и модулей.....	37
5.2 Оценка производительности системы при работе с различными объемами данных.....	37
5.3 Сравнение полученных результатов с ожидаемыми и требованиями.....	39
6. Заключение.....	39
6.1 Подведение итогов исследования.....	40
6.2 Выводы по разработке и реализации системы.....	40
6.3 Рекомендации по дальнейшему развитию и использованию.....	41
7. Список использованных источников.....	42
7.1 Литература, статьи, онлайн-ресурсы, использованные при разработке диплома..	43

8. Приложения.....	44
8.1 Исходный код разработанных скриптов и приложений.....	44
8.2 Примеры данных и результатов анализа.....	45
8.3 Дополнительные материалы, подтверждающие работоспособность и эффективность системы.....	45

1. Введение.

В современном мире объемы данных, с которыми приходится работать в бизнесе, постоянно растут. Каждый день компании генерируют огромные объемы информации, включая данные о клиентах, продажах, финансах, рынке и многом другом. Эффективный анализ этих данных становится ключевым фактором успешного ведения бизнеса. Он позволяет выявлять тенденции, понимать потребности клиентов, оптимизировать производственные процессы, принимать обоснованные решения и многое другое.

Анализ данных в бизнесе — это процесс преобразования сырых данных в полезную информацию с целью выявления закономерностей, трендов и паттернов, которые могут помочь в принятии стратегических и оперативных решений. Он включает в себя различные методы и техники, начиная от простого агрегирования данных и заканчивая сложными моделями машинного обучения.

Однако с ростом объемов данных и сложности аналитических задач становится все труднее реализовать эффективный анализ данных с помощью традиционных методов. В этом контексте становится актуальной задача разработки и внедрения автоматизированных систем анализа данных, которые могли бы обрабатывать большие объемы информации, проводить сложные вычисления и предоставлять пользователю удобный интерфейс для работы с результатами анализа.

В данном дипломном проекте предлагается реализация автоматизированной системы для анализа данных бизнеса с использованием языка программирования Python. Python выбран не случайно — это мощный инструмент для работы с данными, который широко используется в научных и бизнес-сферах благодаря своей простоте, гибкости и богатой экосистеме библиотек для анализа данных.

Целью данного исследования является разработка и реализация системы, способной эффективно обрабатывать большие объемы данных, проводить различные виды анализа и визуализации результатов. Для достижения этой цели поставлены следующие задачи:

1. Разработать архитектуру системы, определить основные компоненты и их взаимодействие.
2. Создать скрипты и приложения для сбора, обработки и анализа данных.
3. Разработать алгоритмы для работы с большими объемами данных.
4. Реализовать функционал визуализации результатов анализа.
5. Провести тестирование системы и оценить ее производительность при работе с различными объемами данных.
6. Сравнить полученные результаты с ожидаемыми и требованиями.

Далее в работе будет представлен обзор существующих решений в области анализа данных бизнеса, постановка задачи и проектирование системы, описание этапов разработки, тестирование и оценка качества системы, а также заключение и рекомендации по дальнейшему развитию проекта.

Работа над данной темой актуальна и значима, так как современным компаниям требуются инструменты, способные обрабатывать и анализировать огромные объемы данных, чтобы принимать обоснованные решения и оставаться конкурентоспособными на рынке. Надеюсь, что результаты данного исследования окажутся полезными и помогут в решении актуальных задач бизнеса.

использованием Python, что поможет повысить эффективность принятия решений и достижение бизнес-целей.

7. Список использованных источников

1. McKinney, Wes. "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython." O'Reilly Media, 2017.
2. VanderPlas, Jake. "Python Data Science Handbook: Essential Tools for Working with Data." O'Reilly Media, 2016.
3. Grus, Joel. "Data Science from Scratch: First Principles with Python." O'Reilly Media, 2015.
4. Raschka, Sebastian, and Mirjalili, Vahid. "Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow." Packt Publishing, 2019.
5. Bengfort, Benjamin, Bilbro, Rebecca, and Ojeda, Tony. "Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning." O'Reilly Media, 2018.
6. Hunter, John D. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
7. Hunter, John D. "Matplotlib: Visualization with Python." Oreilly & Associates Incorporated, 2007.
8. Seabold, Skipper, and Perktold, Josef. "Statsmodels: Econometric and Statistical Modeling with Python." Proceedings of the 9th Python in Science Conference, vol. 57, pp. 61-66, 2010.
9. Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
10. Rossum, Guido van. "Python Programming Language." USENIX Annual Technical Conference, 2007.
11. Python Software Foundation. "Python Language Reference, version 3.9." Python.org, 2020.
12. McKinney, Wes. "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython." O'Reilly Media, 2017.

7.1 Литература, статьи, онлайн-ресурсы, использованные при разработке диплома

1. Python.org - Официальный сайт языка программирования Python, <https://www.python.org/>.
2. Документация библиотеки Pandas - Официальная документация библиотеки Pandas для работы с данными, <https://pandas.pydata.org/docs/>.
3. Документация библиотеки NumPy - Официальная документация библиотеки NumPy для работы с массивами и матрицами, <https://numpy.org/doc/>.
4. Документация библиотеки Matplotlib - Официальная документация библиотеки Matplotlib для визуализации данных, <https://matplotlib.org/stable/contents.html>.
5. Документация библиотеки Scikit-learn - Официальная документация библиотеки Scikit-learn для машинного обучения, <https://scikit-learn.org/stable/documentation.html>.
6. Stack Overflow - Онлайн-платформа для вопросов и ответов по программированию, <https://stackoverflow.com/>.
7. Towards Data Science - Онлайн-платформа с статьями и руководствами по анализу данных и машинному обучению, <https://towardsdatascience.com/>.
8. GitHub - Онлайн-ресурс для хранения и совместной работы над кодом, <https://github.com/>.
9. DataCamp - Онлайн-платформа для обучения анализу данных и программированию на Python, <https://www.datacamp.com/>.
10. O'Reilly Media - Издательство, предоставляющее книги и руководства по программированию и анализу данных, <https://www.oreilly.com/>.