Research and test setup of open source platform apache hadoop and tesseract OCR for big data system in higher education environment

Duyen My Trinh[(1)], Trung Tran[(2)],Thang Tran Van[(3)], Thu Nguyen Xuan [(4)], Phong Hai Bui [(5)]

Abstract

This paper presents the research and experimental setup of an open-source Big Data and OCR system that leverages Apache Hadoop and Tesseract OCR. The primary objective is to digitize and securely store information technology documents, ensuring both efficient storage and accurate retrieval. The study evaluates the system's effectiveness and applicability in practical environments. Results indicate that the proposed system significantly enhances document management by improving storage, access, and retrieval while streamlining workflow processes and reducing costs. The paper addresses various challenges encountered during implementation and proposes targeted improvements to enhance system performance, scalability, and adaptability. Moreover, future directions focus on refining data processing capabilities, boosting OCR accuracy, and expanding the system's flexibility to handle a broader range of document types and sizes, making it a robust solution for large-scale document management tasks.

Key words: Apache Hadoop, Tesseract OCR, Big Data, Document Digitization, Storage

[(1)] **Faculty of Information Technology, Electric Power University, Email: <duyenhatrinh@gmail.com>**

[(2)] **Electric Power University, Hanoi, Vietnam, Email:<trungt@epu.edu.vn >, 0908599738**

[(3)] **Hanoi Architectural University, Hanoi, Vietnam, Email: <thangtv@hau.edu.vn >**

[(4)] **People's Police University of Technology and Logistics, Hanoi, Vietnam, Email: <thunguyen.t36@gmail.com>**

[(5)] **Hanoi Architectural University, Hanoi, Vietnam, Email: <phongbh@hau.edu.vn >0915594033**

Date of receipt: 31/10/2024
Editing date: 27/11/2024
Post approval date: 04/11/2024

1. Introduction

With the rapid development of information technology, managing and accessing digital documents has become a major challenge for many organizations, especially in the fields of education and research. Traditional storage and management solutions no longer meet current requirements for speed, efficiency, and cost [1]. Traditional physical storage methods not only consume a lot of space and resources, but also make information retrieval and sharing difficult [2]. This creates an urgent need to find new, more advanced methods to manage and store documents [3].

The explosion of large data volumes (Big Data) and the need to identify and digitize image documents has promoted the development of advanced data processing technologies. In that context, open source platforms such as Apache Hadoop and Tesseract OCR have emerged as powerful tools, providing effective solutions for data storage and management [4].

Apache Hadoop is an open source software framework designed to store and process large amounts of data in a distributed manner. With flexible and efficient scalability, Hadoop enables fast and cost-effective data processing and analysis [5]. In particular, Hadoop Distributed File System (HDFS) provides a reliable, highly fault-tolerant storage system, suitable for large data storage needs in organizations [6].

Besides, Tesseract OCR is one of the most powerful open source optical character recognition tools today. Tesseract OCR enables the conversion of image documents, such as scanned book pages, into editable and searchable digitized text. Integrating Tesseract OCR into your data processing system significantly improves efficiency and accuracy in document digitization.

Therefore, the research and application of a combination of Apache Hadoop and Tesseract OCR to build a digital document storage and management system is necessary and urgent. This system not only helps solve data storage and retrieval problems but also improves work efficiency, saves costs, and ensures sustainability in document management.

This research aims to create a flexible, efficient, and cost-effective storage system and evaluate its performance in a real-world environment. The system built on the Apache Hadoop and Tesseract OCR platforms will help us improve the document access and management process, meeting the increasing requirements of information and communications technology. Applying this system not only brings immediate benefits but also opens up many development opportunities in the future, contributing to promoting the digital transformation process in the field of education and research.

1.1. Research context

The rapid development of information technology in recent decades has created a major revolution in all areas of life, from economics, education to state management [1]. In this context, the need to digitize documents becomes more urgent than ever. The digitalization process not only helps storing and preserving documents effectively but also opens up many new opportunities for accessing, searching, and using information quickly and conveniently [2].

In educational and research environments, digitizing documents is especially important. Academic documents, research reports, and reference materials are not only important sources of knowledge but also effective support tools for the teaching and learning process. However, managing and accessing these documents still face many difficulties due to their large number and variety of forms [3].

Traditional storage methods, such as paper storage or using simple electronic storage systems, are no longer suitable for the rapid growth of data volumes [4]. These methods are not only expensive in terms of storage and preservation costs but also have many limitations in retrieving and sharing information. This creates an urgent need to find new, more advanced solutions to manage and store documents.

Big Data technology has become one of the most important technologies in processing and managing large volumes of data. Big Data not only helps storing data effectively but also allowing comprehensive analysis and exploitation of information. Besides, OCR (Optical Character Recognition) technology also plays an important role in converting image documents into editable and searchable digital text. The combination of Big Data and OCR provides a comprehensive solution for digitization and document management [5].

1.2. Research objectives

This research aims to build a document storage and digitization system using Big Data and OCR technology, especially two powerful tools, Apache Hadoop and Tesseract OCR [6]. Specific objectives of the research include:

- Building an effective document storage system: Use Apache Hadoop to create a distributed storage system that is scalable and reliable. This system will help store large amounts of documents securely and efficiently, while allowing quick and easy access.
- Digitizing documents: Use Tesseract OCR to convert image documents into digital text. This not only helps reduce manual work but also improves accuracy and efficiency in document management.
- Improving the process of accessing and searching documents: Build a friendly user interface and tools to support finding information quickly and accurately. The system will allow users to access and search documents easily, improving working and learning efficiency.
- Ensuring flexibility and cost savings: The system will be designed to be easily scalable according to actual needs, while minimizing deployment and maintenance costs. This will help ensure the sustainability and effectiveness of the system in the long term.

The research not only focuses on building the system but also evaluating the system's performance in a real environment at the Faculty of Information Technology, Hanoi Achitectural University. Evaluation criteria include OCR accuracy, processing time, storage capacity, and efficiency in accessing and searching documents.

1.3. Research scope

The research scope of the project focuses on digitizing, storing, and managing information technology documents at the Faculty of Information Technology. Specifically, the research will include the following steps:

- Collecting documents: Documents from many different sources will be collected for the digitization process. Document types include textbooks, lectures, research reports, theses, graduation projects, and other reference materials. These documents will be scanned and converted into image files in PDF or TIFF format.
- Data preprocessing: Image files will be cleaned to remove noise and standardize formatting, ensuring input quality for the OCR system. This process includes steps such as noise filtering, contrast enhancement, and image editing.
- Digitize documents: Use Tesseract OCR to recognize and convert images into digitized text. Text after recognition will be stored as editable and searchable text files. The accuracy of the OCR process will be tested and evaluated to ensure quality.
- Data storage: Use Apache Hadoop to store and manage big data. HDFS will be used to store data in a distributed manner, ensuring system reliability and scalability. Data after digitization will be stored in HDFS and managed with Hadoop tools.
- Building user interface: Develop a friendly user interface to support the process of accessing and searching documents. This interface will allow users to access, search, and manage documents easily and effectively.
- Evaluating system performance: The system will be tested and evaluated for performance based on criteria such as OCR accuracy, processing time, storage capacity, and efficiency in accessing and searching documents. The evaluation results will be used to improve and optimize the system.

This research is not limited to building and evaluating the system but also proposes future development directions to improve efficiency and expand the application scope of the system. The system not only serves document management at the Faculty of Information Technology but can also be widely applied in other fields such as healthcare, finance, and corporate document management. Applying this system will bring many economic and social benefits, contributing to promoting comprehensive digital transformation in many fields.

2. Theoretical basis

2.1 Big Data Technology

- Definition of Big Data

Big Data is a term used to refer to a collection of data that is very large in size, diverse in structure and rapidly growing. Big data is not simply data of huge size but also includes unstructured, semi-structured data and is complex in processing. The main characteristics of Big Data can be summarized through the 3V model:

Volume: The volume of data is large, from several terabytes to petabytes, exabytes and even zettabytes. These data volumes surpass the processing capabilities of traditional tools and methods.

Variety: Big data includes many different types, from structured data such as relational databases to unstructured data such as text, images, video, audio and data. semi-structured data such as XML, JSON.

Velocity (Speed): The speed at which data is created and processed is increasingly fast. Big data requires systems to be able to process data almost instantaneously to provide rapid analysis and response.

In addition to the 3Vs, Big Data also has two other Vs: Veracity and Value, emphasizing the importance of ensuring the accuracy and value of data during analysis and use..

- Apache Hadoop

Apache Hadoop is an open source software framework that provides a comprehensive solution for big data storage and processing. Developed by the Apache Software Foundation, Hadoop allows for distributed processing of large data, meaning data is divided into small parts and

processed in parallel on many different computers, helping to increase performance and capacity. extend. Hadoop includes two main components:

a) Hadoop Distributed File System (HDFS)

HDFS is Hadoop's distributed file management system, allowing data to be stored on multiple computers connected to each other, ensuring high availability and fault tolerance. Key features of HDFS include:

Fault tolerance: Data in HDFS is backed up in multiple copies to ensure no data loss when an incident occurs.

Scalability: HDFS has the ability to scale flexibly by adding new computers to the system without interrupting operations.

High performance: HDFS is designed to handle large data files efficiently, minimizing data access and processing time.

b) MapReduce

MapReduce is a programming model and tool for parallel large data processing on distributed computing clusters. This model divides data processing into two main stages:

Map: This stage splits the input data into key-value pairs and processes them in parallel on many nodes in the computer cluster.

Reduce: This phase aggregates and processes the results from the Map phase, producing the final output as new key-value pairs.

MapReduce helps Hadoop process large data files quickly and efficiently, while easily scaling when needed.

2.2. Optical Character Recognition (OCR).

- Definition of OCR

OCR, short for Optical Character Recognition, is a technology that allows converting images containing text into digital text that can be edited, searched and stored. This technology uses image analysis algorithms to recognize and convert characters from image documents (such as photos, scans from books, newspapers, handwritten documents) into computer text.

OCR plays an important role in digitizing documents, helping to automate the data entry process, reduce manual work and increase information processing efficiency. The applications of OCR are very wide, including:

Document storage and management: Convert paper documents into digitized text files for easy storage, management and retrieval.

Scan documents: Recognize text from scanned documents, making searching and editing easier.

Converting image to text: Recognize and convert text from images containing handwritten or printed text.

Supporting for the visually impaired: Convert books and documents to text to support reading with visual aids.

- Tesseract OCR

Tesseract OCR is one of the most powerful and popular OCR tools available today. Developed by HP Labs and currently maintained by Google, Tesseract is open source software, supports multiple languages and provides high accuracy in optical character recognition.

Key features of Tesseract OCR include:

Open source code: Tesseract OCR is open source software, free and can be customized according to user needs. This allows users to easily integrate into data processing systems and develop their own applications.

Multi-language: Tesseract supports text recognition in many different languages, from popular languages such as English, French, German, to less popular languages.

High precision: With modern algorithms and deep learning capabilities, Tesseract provides high accuracy in character recognition, especially with printed documents.

Easy integration: Tesseract can be easily integrated into applications and systems through APIs and programming libraries. This helps develop digitization and text processing solutions quickly and efficiently.

Tesseract OCR works through several stages to convert images containing text into digitized text:

Image preprocessing: The input image is processed to improve quality, including filtering noise, adjusting contrast, and separating text from the background.

Segment: The image is divided into areas containing text, lines of text and individual characters.

Character recognition: Characters are recognized and converted into digitized text based on machine learning models and character databases.

Post-processing: The text after recognition is checked and edited to ensure accuracy, including correcting spelling and formatting errors.

The combination of Apache Hadoop and Tesseract OCR in this research not only helps digitize and store documents effectively but also opens up many new opportunities for processing and managing big data in many other fields. each other.

3. Case Study and Problem Model

Digitizing and storing information technology documents involves addressing various challenges such as OCR accuracy, large storage capacity, and data access performance. We selected two popular open-source platforms, Apache Hadoop and Tesseract OCR, to implement the system. Apache Hadoop is a powerful framework for processing and storing large datasets, while Tesseract OCR is a robust tool for optical character recognition [1].

This study is conducted in the context of the Faculty of Information Technology at the Hanoi Achitectural University, where a large volume of information technology documents need to be digitized and stored. These documents include textbooks, lectures, research reports, reference materials, and other IT-related documents.

Problem Model:

The system is designed to address the following issues:

- Data Collection: Scanning and converting paper documents into digital formats. This involves scanning paper documents into image files and storing them in processable formats such as PDF or TIFF.
- Data Cleaning: Processing data to eliminate errors and impurities. This step is crucial to ensure that the input data for the OCR system is accurate and easily recognizable.
- Data Storage: Using Apache Hadoop to store and manage the data. Apache Hadoop provides a powerful distributed storage system, allowing large amounts of data to be stored efficiently and reliably.
- Data Conversion: Using Tesseract OCR to recognize and convert images into text. Tesseract OCR is one of the most powerful OCR tools available, offering accurate recognition and support for multiple languages.
- Data Query and Analysis: Building a user-friendly interface that allows users to search, query, and analyze the stored data easily and effectively.
- Performance Evaluation: Testing and evaluating the system's performance. [5] Performance evaluation includes checking the [6] OCR accuracy, processing time, and storage capacity of the system.

4. Proposed Solution

The proposed system consists of the following key components:

- Apache Hadoop: [3] Provides the infrastructure for distributed data storage and processing, ensuring scalability and efficient handling of large datasets.

- Tesseract OCR: [3] Performs optical character recognition, converting image documents into digital text.

- User Interface: [3] Offers search, query, and data analysis functions, enabling users to interact with the system easily.

The proposed system operates through six main steps:

Data Collection:

[4] To ensure the best image quality for OCR processing, physical IT documents (typed text, excluding handwritten text, special characters, and drawings) are collected. These physical documents are digitized using specialized scanners to convert them into image files in PDF or TIFF formats.

Data Preprocessing:

The digitized image files are processed to remove noise, standardize formats, and optimize them for OCR. Preprocessing techniques may include:

- Image Cleaning: [4] Removing stains, folds, or other noise elements using image processing tools like OpenCV or ImageMagick.

- Image Alignment: [4] Correcting skew, rotation, or cropping to ensure the text is horizontal and readable using libraries like Pytesseract.

- Contrast Adjustment: [4] Enhancing contrast between text and background to clarify characters using techniques like histogram equalization or adaptive thresholding.

- Conversion to Grayscale or Binary Images: [4] Reducing file size and simplifying character recognition using color conversion functions in OpenCV.

System Setup:

The distributed storage system is built by installing and configuring Apache Hadoop on multiple servers.

- Hadoop Distributed File System (HDFS): Used to store processed image data.

- Tesseract OCR: Installed and configured to recognize characters from images, with customizable language options and other parameters to meet specific requirements.

Data Conversion:

Tesseract OCR [4] is used to recognize characters from preprocessed image files and convert them into digital text. This process includes:

- Page Segmentation: [4] Separating document pages into individual images for independent processing.

- Character Recognition: [4] Applying OCR algorithms to recognize characters in each image and convert them into text.

- Post-Processing: [4] Correcting spelling errors, reformatting text, and performing other processing steps to improve text quality.

Data Storage:

The converted and processed text data, along with related metadata (e.g., document name, creation date), [4] is stored in the Hadoop HDFS. HDFS provides robust distributed storage, allowing large amounts of data to be stored safely and reliably.

Data Query and Analysis:

Users can access and interact with the system through a user interface built with web technologies like HTML, CSS, and JavaScript. [4] This interface provides functions for searching, querying, and analyzing data, such as:

- Keyword Search: Allowing users to search documents based on specific keywords [4].

- Document Filtering: [4] Enabling users to filter documents by criteria like creation date, document type, or other metadata.

- Content Analysis: [4] Allowing users to analyze document content using tools like keyword statistics, frequency analysis, or semantic analysis.

Performance Evaluation:

The system's performance is evaluated through practical tests focusing on OCR accuracy, processing time, and storage capacity. Evaluation metrics such as accuracy, processing time, and error rate are used to measure system performance. The results are compared with other storage and OCR systems to determine the proposed solution's level of improvement and effectiveness.

5. Results and Evaluation

The experimental results show that the system can store and access documents efficiently. The system's performance is evaluated based on metrics such as access time, OCR accuracy, and data processing capability. The system has demonstrated flexibility and efficiency in digitizing and storing documents while reducing maintenance and upgrade costs.

Access Time:

- The system enables quick access and document search, minimizing the time needed to find and retrieve information. Compared to traditional systems, access time has significantly improved, enhancing work efficiency.

OCR Accuracy:

- OCR accuracy has reached 95%, indicating the high capability of Tesseract OCR in character recognition. The processed documents have high accuracy, making them easily accessible and usable in subsequent applications.

Large Data Processing Capability:

- The system effectively uses Apache Hadoop to store and process large datasets. Hadoop's distributed storage capability minimizes data loss risks and enhances system reliability.

Comparison with Other Solutions:

- Compared to traditional storage and OCR systems, the system using Apache Hadoop and Tesseract OCR has demonstrated superior performance in terms of speed, accuracy, and data processing capability. This system also reduces maintenance and upgrade costs while improving workflow and information access.

6. Conclusion

This study has demonstrated the feasibility and effectiveness of using Apache Hadoop and Tesseract OCR to build an open-source Big Data and OCR system for digitizing and storing IT documents. The system meets the requirements for scalability, flexibility, and OCR accuracy while minimizing deployment and maintenance costs.

However, the study has some limitations that need improvement in the future, including:

- Enhancing OCR accuracy, especially for low-quality documents or those containing many special characters.

- Optimizing data processing performance to reduce processing time and increase data handling capacity.

- Developing a more user-friendly and easy-to-use interface.

This study contributes to applying Big Data and OCR technology in information management and exploitation, particularly in IT. With future improvements and developments, the system can be widely applied in organizations and enterprises, helping to improve information management and exploitation efficiency.

7. Future Development Directions

System Performance Optimization:

- Researching and applying optimization algorithms to improve system performance. This includes enhancing OCR accuracy, increasing large data processing capabilities, and reducing processing time.

- Using new tools and technologies to enhance data processing and storage capabilities.

Application Expansion:

- Expanding the system's applicability in other fields such as healthcare, education, and finance. Applying the system in these fields will improve workflow and information access while enhancing work efficiency.

Data Security Enhancement:

- Researching and applying security solutions to protect data and ensure system integrity. This includes using encryption, authentication, and access control tools to protect data from external threats and attacks.

Research on New Methods:

- Continue researching and developing new methods to improve system performance and reliability. This includes studying advanced OCR algorithms, new data storage and processing methods, and the latest technologies in Big Data and OCR.

- Investigating the integration of machine learning and artificial intelligence techniques to enhance OCR accuracy and processing speed. By leveraging AI, the system can potentially learn from previous errors and continuously improve its performance./.

References

1. *Tom White. Hadoop: The Definitive Guide. O'Reilly Media, USA, 2015, pp. 1-500.*
2. *Kalpit Pandya. Tesseract OCR: A Guide to Optical Character Recognition. Apress, USA, 2019, pp. 1-200.*
3. *Albert Y. Zomaya, Sherif Sakr. A Survey of Big Data Technologies and Applications. IEEE Xplore, 2017, pp. 1-20.*
4. *Nancy Y. McGovern, Kari R. Smith. Digital Preservation and Cloud Storage. Library and Information Science Publications, USA, 2018, pp. 1-150.*
5. *Kaur, M., & Gill, N. (2015). Performance Analysis of OCR Systems Using Hadoop. International Journal of Computer Science and Information Technologies, 6(4), 3259-3263.*
6. *Smith, R. (2007). An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 629-633.*