Benefits and Security Challenges of Big Data Analytics

Alexander I. Iliev [0000-0002-4220-3637]

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
8, G. Bonchev Str., 1113, Sofia, Bulgaria
SRH Berlin University of Applied Sciences, 10, Ernst-Reuter-Platz, 10587, Berlin, Germany
ailiev@berkeley.edu

Abstract. Big data analytics is a powerful technique used by companies to secure their data. However, data security remains a significant challenge for cultural entities such as libraries and museums, as hackers can target database systems and gain access to sensitive information. This study examines the benefits and security issues associated with big data analytics and proposes a manual approach for analyzing data with application to digital presentation and preservation of cultural heritage. By providing insights into the strengths and weaknesses of big data analytics and exploring strategies for securing data, the goal of the research is to help organizations make informed decisions about their data management practices.

Keywords: Data Security, Data analytics, Cultural and Scientific Heritage.

1 Introduction

In this digital era, everyone is increasingly accessing internet servers and online communication platforms to collect vast amounts of data from various users. However, managing and making sense of this data can be a daunting task, especially when it comes to preserving items with historical value. Culture and heritage are fundamental aspects of society that define a community's identity and traditions.

Effectively managing and leveraging data is essential for the cultural and scientific community to broaden their horizons and expand their impact. This encompasses various data types, including valuable cultural artifacts, historical records, scientific research findings, technical specifications for specific databases related to art and historical data, and even rich artistic data. Analyzing this data can yield significant insights that positively impact strategies and services, enabling the community to enhance their understanding of culture, art, history, heritage, and science and cater to the needs of researchers, enthusiasts, and the public more effectively. Furthermore, it plays a vital role in the digital presentation and preservation of cultural heritage, facilitating the exploration and dissemination of invaluable cultural artifacts and scientific knowledge to a wider audience, thereby enriching our collective understanding and appreciation of our cultural heritage and scientific advancements.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 13, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2023. ISSN: 1314-4006, eISSN: 2535-0366

Big data analytics is a leading technology that provides a platform to manage data handling problems and enables them to review the complex data obtained from the internet, including cultural data. Using big data analytics is crucial for many building companies, which helps them outperform during the age of competition.

Bigger analytics can lead to better trial results and help predict better outcomes by analyzing cultural data alongside other data types. In terms of decision making, big data analysis plays a significant contribution in proposing important conclusions by studying data from cultural sources.

However, securing data is a challenging task related to big data analytics, especially when it comes to preserving cultural heritage. It's crucial to ensure that the data is collected and analyzed for the purpose of identifying and managing unwanted signals posed by attackers. The purpose of this study is to review and evaluate the benefits and security risks linked with big data analytics, with a specific focus on preserving cultural heritage and values. By examining the strengths and weaknesses of this technology, we aim to provide insights that can help organizations make informed decisions about cyber security. We will explore best practices for securing big data, including cultural data, and identify potential threats to its security. By balancing the benefits and challenges of big data analytics with the preservation of cultural heritage and values, organizations can maximize the value of big data while safeguarding their cultural heritage sources.

1.1 Background and Related Work

In today's digital age we are increasingly generating and collecting vast amounts of data from internet servers, including structured and unstructured data. Big data analytics offers significant tools and programs to effectively review and analyze such data sets. Wang et al. (2018) reported that big data analytics has the potential to extract valuable information from complex datasets, enabling the discovery of human behavior related to the use of cultural datasets. Analytical systems can uncover hidden patterns and insights related to the ways in which we use data related to art, history, and culture.

This study aims to deliver a more in-depth examination of the role of cyber security as applied to heritage and culture. The use of big data analytics has a significant impact on heritage and culture, especially in various fields of the arts, and history, where it can not only help to preserve cultural heritage, but also to facilitate cultural exchange.

To address research gaps from previous studies, we will conduct a systematic literature review to examine the security issues and threats linked with big data analytics (Babak Bashari Rad, 2016). This will enable us to provide insights into best practices for securing big data and identify potential threats to its security.

It is worth noting that the origins of data science can be traced back to the major project initiated in 1937 by Franklin D. Roosevelt's administration in the USA, which aimed to keep track of contributions from 26 million Americans. Today, there has been a significant increase in big data startups, all trying to deal with big data and help other big organizations elaborate on big data. More and more companies are slowly adopting big data as a solution, which has the potential to impact not only the business world but also the heritage and culture sectors.

1.2 Benefits of Big Data Analytics

Described below are several benefits of utilizing big data analytics:

- Market Knowledge
- Recruitment of Employees
- Improved Strategy and Planning
- Increased efficiency
- Controls Online Reputation
- Time reduction
- New Product Development
- Increased sales
- Cost Saving
- Focus on client's needs.

1.3 Cost Optimization

Big data has emerged as a crucial technology for companies looking to manage their data handling issues. One of the major advantages of big data analytics is that it offers cost benefits to companies by providing a single system for storing, processing, and analyzing large data sets. By implementing a single system, cultural entities can avoid the need for additional networks and sources, making the data collection process more efficient.

As noted by Zhao and Ke (2019), big data analytics can streamline the data handling process by allowing cultural institutions such as museums and libraries to collect data from the internet, process complex data, and perform data analysis activities all within one system. This not only saves time and money but also allows these parties to make more informed decisions based on insights derived from the data.

One example of cost optimization in cultural heritage preservation is the implementation of digital preservation techniques. By digitizing cultural artifacts, such as ancient manuscripts, paintings, or archaeological findings, organizations can reduce the costs associated with physical storage, maintenance, and conservation. Digitized collections can be stored in secure digital repositories, eliminating the need for extensive physical infrastructure. Additionally, digital preservation allows for easier accessibility and sharing of cultural heritage, enabling broader public engagement, and reducing costs related to physical exhibition spaces and transportation. Through cost optimization strategies like digital preservation, organizations can efficiently allocate resources, ensure long-term preservation of cultural heritage, and make it more accessible to a global audience.

Furthermore, big data analytics can help museums and libraries preserve collections of cultural heritage by allowing for more accurate and precise data collection and analysis. This can be particularly important in industries where cultural sensitivity and understanding are essential, such as the tourism industry.

2 Popular types of Security Treads related to Big Data Analytics

2.1 Malware Attacks

Malware attacks are a significant security threat to big data networks and systems, as they can compromise the privacy and security of data (Deli et al., 2017). These attacks involve the use of malicious codes that can transfer unwanted signals to specific data networks that deal with sciences, arts or historic artifacts for example, allowing hackers to access stored data and perform other hacking-related operations. Unfortunately, big data networks are particularly vulnerable to these attacks, as they can produce larger viruses and worms that target database and storage systems used by companies. This can lead to significant security challenges (Aditya & Youddha, 2016) as big data networks are often unable to detect and find malware signals in the system. As a result, companies are facing an increasing number of cyber-attacks each year, with the total number of malware infections rising from around 300,000,000 in 2016 to around 8000,000,000 in 2020. To protect themselves against these threats, cultural centers must take a comprehensive approach to security that includes measures such as encryption, multi-factor authentication, and network monitoring. By implementing these strategies, cultural entities can minimize their risk of malware attacks and ensure the security of their data.

Next is a simple example of how to prevent a malware attack using Python:

```
import os

def is_valid_file_extension(file_name):
    """
    Check if the file extension is valid or not.
    """
    valid_extensions = ['.csv', '.xlsx', '.docx', '.pdf']
    file_extension = os.path.splitext(file_name)[1].lower()
    return file_extension in valid_extensions

# Example usage
file_name = 'example.exe'
if not is_valid_file_extension(file_name):
    print(f'Error: {file_name} has an invalid file extension')
else:
    print(f'{file_name} is a valid file')
```

In this example, we define a function `is_valid_file_extension` that takes a file name as input and checks if the file extension is valid or not. We define a list of valid file extensions and use the `os.path.splitext` function to get the file extension. We then convert the file extension to lowercase and check if it is in the list of valid extensions. Finally, we demonstrate the usage of the function by checking if a file named *example.exe* has a valid file extension.

This is just a basic example but validating file extensions is an important step in preventing malware attacks. You can build on this example and add more complex logic to further prevent malware attacks.

2.2 Phishing Threats

Phishing attacks are a significant threat to big data networks, as they can compromise the privacy and security of cultural entities' communication channels and internet servers. These attacks involve the use of fake or fraudulent signals that appear to come from legitimate sources and can be difficult to detect and prevent. In the realm of digital preservation of cultural and scientific heritage, the implementation of robust phishing threat prevention measures is crucial to safeguard valuable and sensitive digital assets. Phishing attacks pose a significant risk to the integrity and security of digital collections, as they target unsuspecting users with deceptive emails or messages aimed at stealing credentials or spreading malware. To protect against these threats, organizations engaged in digital preservation must prioritize employee training on recognizing and reporting phishing attempts. Implementing email filtering systems and robust authentication protocols can further fortify defenses against phishing attacks. Additionally, regular security audits and updates to software systems can help identify vulnerabilities and mitigate potential risks. By actively countering phishing threats, organizations can ensure the preservation and integrity of digital cultural and scientific heritage, maintaining public trust and accessibility to these valuable resources.

Here is a Python example of a simple Phishing Attack prevention:

```
import phishingcheck

url = "http://www.example-phishing-site.com"
result = phishingcheck.check(url)

if result.is_phishing:
    print(f"{url} is a phishing website.")
else:
    print(f"{url} is not a phishing website.")
```

In this example, we first import the `phishingcheck` package which provides a simple API for checking URLs against a list of known phishing websites. We then define the `url` variable as the URL we want to check.

Next, we call the `check` method of the `phishingcheck` package, passing in the URL we want to check as an argument. This method returns a `PhishingCheckResult` object which contains information about whether the given URL is a phishing website or not.

Finally, we use an if statement to check if the `is_phishing` attribute of the `PhishingCheckResult` object is True or False, and print out a message indicating whether the URL is a phishing website or not.

2.3 DOS and DDOS Attacks

Denial of service (DoS) attacks are a common security threat to big data networks, as they provide a platform for cyber criminals to target computing networks and systems (Ahmad et al., 2018). These attacks involve the transfer of traffic signals to big data networks, allowing hackers to enter servers and create significant security risks. Distributed denial of service (DDoS) attacks are a particularly insidious form of DoS attack, as they involve the distribution of botnets and malicious codes into the computing

networks used by libraries and museums, making it easier for hackers to gain login credentials and compromise security.

In response to a given DDoS attack, here is a Python code that counters it to prevent it from happening:

```
import socket

# create a socket object
server_socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)

# bind the socket to a public host, and a well-known port
server_socket.bind(('', 80))

# become a server socket
server_socket.listen(5)

# set a maximum client count
max_clients = 10

# define a counter to keep track of number of clients
client_count = 0

while True:
    # accept a client connection
    (client_socket, client_address) = server_socket.accept()

    # increment the client count
    client_count += 1

    # if we exceed the maximum client count, reject the connection and close
    # the socket
    if client_count > max_clients:
        client_socket.close()
        continue

    # otherwise, read the incoming data from the client
    data = client_socket.recv(1024)

    # if the incoming data is empty, the client has disconnected, so decrement
    # the client count and close the socket
    if not data:
        client_count -= 1
        client_socket.close()
        continue

    # print the incoming data to the console
    print(data)
```

In this code, we added a maximum client count of 10. When a new client tries to connect, we increment the client count. If the client count exceeds the maximum count, we reject the connection and close the socket. We also check if the incoming data is empty, indicating that the client has disconnected, and decrement the client count and close the socket accordingly.

By setting a maximum client count, we can prevent DoS attacks that try to overwhelm the server by opening too many connections at once.

The field of digital preservation of cultural and scientific heritage necessitates robust measures to prevent and mitigate DoS and DDoS attacks, which can disrupt access to valuable digital resources. To protect against these attacks, organizations involved in digital preservation must implement network security measures such as firewalls, traffic monitoring systems, and intrusion detection systems. These tools help detect and mitigate abnormal traffic patterns, identify potential attack sources, and prevent the

overwhelming of network resources. Furthermore, organizations can employ load balancing techniques and utilize content delivery networks (CDNs) to distribute traffic efficiently and handle sudden spikes in user demand. Regular vulnerability assessments and security audits are crucial to identify and patch potential weaknesses in the network infrastructure. By implementing strong DoS and DDoS attack prevention strategies, organizations can ensure the uninterrupted availability and accessibility of digital cultural and scientific heritage, preserving these invaluable resources for future generations.

2.4 Ransomware Threats

Ransomware attacks are a particularly insidious form of malware attack that can have a direct impact on the performance and security of cultural big data networks. These attacks involve the transfer of malicious and unauthorized activities that target computing devices and networks used by libraries and museums, allowing cyber criminals to compromise the privacy and security of big data systems. With the help of ransomware attacks, hackers can encrypt sensitive data obtained from internet servers, making it inaccessible to employees without the permission of the attackers. This poses a significant security threat, highlighting the need for robust security measures and employee training to prevent ransomware threats. By implementing strategies such as data backups, encryption, and anti-malware software, cultural centers can minimize the risk of these attacks and ensure the security of their big data networks.

In the field of digital preservation of cultural and scientific heritage, the prevention of ransomware threats is of paramount importance to safeguard valuable digital assets. Ransomware attacks pose a significant risk to the integrity and accessibility of digital collections, as they can encrypt critical data and demand ransom for its release. To protect against such threats, organizations engaged in digital preservation must implement a multi-layered defense strategy. This includes regular data backups stored securely offline or on isolated systems, employing robust antivirus and anti-malware software, and regularly updating and patching software and operating systems to address vulnerabilities. User awareness and training programs are also crucial to educate staff on safe browsing habits, avoiding suspicious email attachments, and practicing good cybersecurity hygiene. By implementing proactive ransomware prevention measures, organizations can ensure the preservation of invaluable cultural and scientific heritage, protecting it from malicious attacks and ensuring its continued accessibility for future generations.

Next, we show a simple example of how to prevent a ransomware attack with Python:

```
import os

def scan_files(folder_path):
    for root, dirs, files in os.walk(folder_path):
        for file in files:
            file_path = os.path.join(root, file)
            if file.endswith('.exe') or file.endswith('.dll'):
                scan_file(file_path)

def scan_file(file_path):
    # Code to scan file for ransomware using antivirus software
    # If ransomware is detected, the file is quarantined or deleted
    # Otherwise, the file is marked as safe

# Example usage
folder_path = 'C:/Users/MyUser/Documents'
scan_files(folder_path)
```

This code defines a `scan_files` function that recursively scans a specified folder and all its subfolders for files with executable extensions (.exe and .dll in this case), and then passes each file to the `scan_file` function for scanning.

The `scan_file` function contains the code to scan each file for ransomware using antivirus software. If ransomware is detected, the file is either quarantined or deleted. If no ransomware is detected, the file is marked as safe.

By regularly running this code to scan files on a computer or network, organizations can proactively prevent ransomware attacks and protect their sensitive data from being encrypted and held for ransom.

3 Proposed Method

This study adopts an inductive approach to formulate a comprehensive risk assessment strategy specifically tailored for the preservation of cultural and scientific heritage. By employing qualitative research methods, the study delves into the intricate security challenges entailed in the utilization of big data analytics within this field. Through qualitative analysis, the study not only obtains valuable theoretical insights but also captures nuanced details that underpin the research, contributing to a robust foundation for addressing the unique security concerns prevalent in cultural and scientific heritage preservation.

This paper proposes 7 step methods for risk assessment as described in figure 1 and depicted in figure 2:

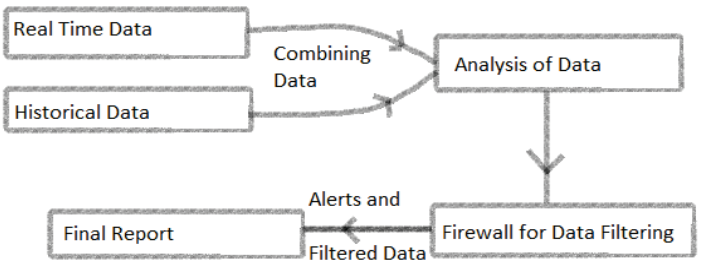

Fig. 1. Method flowchart.

Detailed algorithm steps are shown next:

Step 1: Choose the subset of data from the entire cultural or scientific heritage dataset:

β, μ

Step 2: Select the parameter in each data set.

β = *Real Time Data,* μ = *Historical Data,* π = *Firewall*

Example:

β = *(a, b, c),* μ = *(a-z)*

Step 3: Combining all the dataset related to the attack.

$X = \beta \cup \mu$

Step 4: Analysis of the combined dataset.

$\alpha = \beta \cap \mu$

Step 5: Repeat steps 4 till data is separated.

Step 6: Check the Firewall.

Step 7: Final Report is generated.

In the context of cultural and scientific heritage preservation, the proposed system initiates by capturing real-time data from the operational ecosystem. Acquiring this data involves employing a descriptive analysis process that facilitates efficient and comprehensive summarization of information. To ensure a holistic perspective, it becomes imperative to gather not only current data, but also historical data (data collected from past performance of specific devices) associated with similar systems or contexts. Once the real-time and historical data are obtained, the subsequent step involves their integration and comparative analysis. Real-time data plays a pivotal role in identifying prevailing vulnerabilities and external security threats, enabling proactive measures to detect and thwart potential cyber-attacks. By combining this data with historical insights, the system can identify patterns, simplifying the determination of network security solutions. In certain cases, valuable solutions may already exist, leveraging input from knowledgeable administrators. Overall, this approach optimizes the identification process and fosters the implementation of robust and tailored security measures for the preservation of cultural heritage.

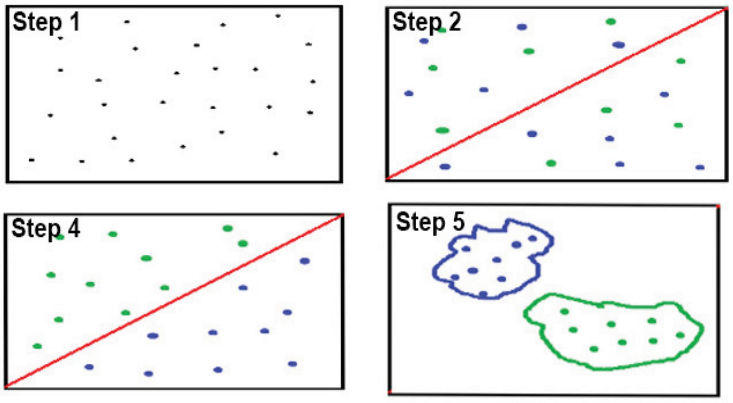

Fig. 2. Graphical representations of steps 1, 2, 4, 5.

Once the data relevant to cultural and scientific heritage preservation have been acquired and amalgamated, the subsequent phase entails meticulous analysis to unveil any emerging or potential security threats. This involves scrutinizing past data in correlation with ongoing activities, enabling the system to identify any anomalous behavior or outliers that may signify a potential attack. By representing the data in a statistical format and employing firewalls, the administrative team gains enhanced capabilities to discern which incoming traffic warrants permission, and which should be restricted. The statistical representation facilitates accessibility for individuals at various proficiency levels, empowering even inexperienced personnel to make well-informed decisions. The inclusion of past data assumes a vital role in this process, providing experts with profound insights into the nature of the attack and its interrelation with other irregularities. By harnessing the power of historical data, the system can effectively discern potential threats and implement robust security measures to prevent them, thereby fortifying the preservation of cultural and scientific heritage.

The final step in this method involves workflow automation, which refers to the execution of the final report. The automated workflow plays a crucial role in identifying potential fraud and reporting it to the security team. By submitting any suspicious events for further investigation in the form of a prepared report, the system can help ensure that any security threats are promptly addressed. The report provides a comprehensive overview of the data that has been collected and examined, highlighting any potential loopholes or vulnerabilities in the organization's data. This automated workflow helps to streamline the process of identifying and addressing security threats, enabling organizations dealing with preservation of cultural and scientific heritage to protect their data and prevent fraud more effectively.

4 Discussion and Recommendations

The proposed method for preventing cyber-attacks, as described in the text, can be effectively applied to preserving cultural heritage datasets. In order to protect cultural

heritage datasets from cyber-attacks, it is important to first acquire real-time data from the working system, as well as past data related to the same or similar systems. This data can then be combined and compared to identify any potential security threats or abnormal behavior. By utilizing descriptive analysis and statistical representations of the data, inexperienced personnel can also participate in the identification process and make informed decisions.

In addition to identifying potential security threats, the proposed method also involves workflow automation for reporting and addressing any suspicious events. This is particularly important for cultural heritage datasets, as they contain sensitive and valuable information that must be protected from cyber-attacks and fraud. By automating the workflow and providing a prepared report, security teams can promptly investigate and address any security threats to the cultural heritage datasets. This approach can help to ensure that cultural heritage is preserved and protected for future generations, while also leveraging the benefits of big data analytics for more effective security solutions.

5 Conclusions and Future Work

In the realm of cultural and scientific heritage preservation, the proposed system presents a unified platform that enables real-time data analysis and comprehensive information retrieval. Implementation of encryption techniques and firewall tools assumes paramount importance to fortify the safety and privacy of the invaluable cultural and scientific heritage data. By employing encryption, the system ensures that data remains protected from unauthorized access or interception. Simultaneously, the utilization of firewall tools bolsters the prevention of malicious intrusions and unauthorized network access, reinforcing the overall security framework. Together, these security measures enable the proposed system to uphold the preservation and safeguarding of cultural and scientific heritage, ensuring its enduring integrity and restricted accessibility.

References

Aditya D. M., & Youddha B. S. (2016). Big data analytics: Security and privacy challenges. In *2016 International Conference on Computing, Communication & Automation (ICCCA)* (pp. 50-53). https://doi.org/10.1109/CCAA.2016.7813688

Ahmad, S., Yasin A., & Shafi, Q. (2018). DDoS attacks analysis in bigdata (hadoop) environment. In *2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (pp. 495-501). https://doi.org/10.1109/IBCAST.2018.8312270

Deli, M. S. M., Ismail, S. A., Kama, N., Yusop, O. M., Azmi, A., & Yahya, Y. (2017). Malware log files for Internet investigation using hadoop: A review. In *2017 IEEE Conference on Big Data and Analytics (ICBDA)* (pp. 87-92), https://doi.org/10.1109/ICBDAA.2017.8284112

Elgendy, N., & Elragal, A. (2014). Big data analytics: a literature review paper. In Perner, P. (eds), *Advances in Data Mining. Applications and Theoretical Aspects.*

ICDM 2014. Lecture Notes in Computer Science (LNAI, volume 8557) (pp. 214–227). Springer, Cham. https://doi.org/10.1007/978-3-319-08976-8_16

Wang, Y., Kung, LA., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3-13. https://doi.org/10.1016/j.techfore.2015.12.019

Zhao, Ch., & Ke, X. (2019). Discussion on the Introduction of the Third Basic Method of Management Accounting from the Perspective of Big Data. In *Proceedings of the 2019 4th International Conference on Modern Management, Education Technology and Social Science (MMETSS 2019)* (pp. 740-744). https://doi.org/10.2991/mmetss-19.2019.149

Received: March 30, 2023
Reviewed: April 22, 2023
Finally Accepted: May 29, 2023