BIG DATA SECURITY FOR AUTOMATIC TEMPER DETECTION

Mr S.R.tribhuwan, Pathare Rushikesh, Khemnar Jagdish, Sonawane Gaurav, Boarde Chaitanya

Lecturer Cloud Computing and Big Data, Padmashri Dr Vitthalrao Vikhe Patil Institute of Technology and engineering (polytechnic), Maharashtra, India
Student Cloud Computing and Big Data, Padmashri Dr Vitthalrao Vikhe Patil Institute of Technology and engineering (polytechnic), Maharashtra, India
Student Cloud Computing and Big Data, Padmashri Dr Vitthalrao Vikhe Patil Institute of Technology and engineering (polytechnic), Maharashtra, India
Student Cloud Computing and Big Data, Padmashri Dr Vitthalrao Vikhe Patil Institute of Technology and engineering (polytechnic), Maharashtra, India
Student Cloud Computing and Big Data, Padmashri Dr Vitthalrao Vikhe Patil Institute of Technology and engineering (polytechnic), Maharashtra, India

ABSTRACT

Big data as well as other forms of digital resources have become more popular due to technical developments in recent times. Massive volumes of data are produced by big data, which may then be analyzed and put into practice in a wide range of technological and scientific fields. While big data has numerous practical uses, it also presents a number of obstacles in the areas of statistics, administration, and confidentiality and safety that must be overcome to enhance the customer experience. The big data is highly susceptible to increasing attacks and other types of manipulation by malicious users. This is highly problematic as the big data is very large and difficult to maintain and preserve to achieve effective understanding of the data that can be useful in achieving insightful information and knowledge. Improving the security of the big data can also lead to decrease in performance and a considerable delay in the processing of the queries on such data, therefore, there is a need for an effective mechanism that can increase the efficiency and the security of the big data at the same time. Therefore, this approach utilizes effective parallel computation for Map Reduce operations as well as forensic analysis for tamper detection through the use of bilinear pairing and recognition of the avalanche effect for report generation. The approach has been effectively tested for the purpose of understanding the effectiveness which has resulted in satisfactory outcomes.

Keywords: : *Big data, Parallel Computation, Bilinear Pairing, Avalanche effect, Data Mapping.*

INTRODUCTION

A growing number of governmental and commercial organizations, including federal agencies, banks, hospitals, and health insurers, have made steps to render their information accessible online in current history. That is, such businesses have been mining the personal information of their customers and users for profit. The digital output volume in many cases is big data, measured in terabytes, and encompasses large and complicated data. Several scholars have lately likened big data to crude oil in regards to the effect it will have on civilization. These datasets are notoriously difficult to work with because to their complexity and lack of guidance, and they come from a wide variety of different places, including transactional databases, Sensing technologies, social networking sites, medical databases, and picture and video repositories, to name just a few.

In order to characterize, understand, or identify intriguing trends, the process of harvesting trends from massive data sets is carried out. Data mining is a multifaceted area within computer engineering that refers to this method. Information retrieval from data' is yet another phrase for the end result of data mining, and the two concepts are often used interchangeably. It is via these discovered and extracted connections that data mining techniques

function. Data mining routinely use inductive techniques and information processing techniques, both of which are included here. Still, there are a few instances when data mining is little more than a rudimentary step in the investigative process.

The amount of information in existence is rapidly expanding. As time goes on, the amount of data created on a worldwide scale is predicted to increase at an exponential rate. The big data ecosystem is one in which data collection and analysis occur on a continuous basis. It is common practice for businesses and other organizations to utilize the information they gather to tailor their offerings to individual customers, enhance the efficiency with which they make decisions, anticipate market shifts, etc. In today's business world, data plays a key role. The cloud or a virtualization software are common places for storing huge data. To complete a job in shared storage, it takes the combined efforts of numerous nodes. In this way, the accuracy of computations may be compromised by attacks on a single or several sites. The storage systems must shoulder a much heavier load to ensure data is secure when it is stored in a distributed fashion. Authentication process becomes more involved if encrypted data is stored. For this reason, it is challenging to directly employ conventional both asymmetric and symmetric encryption algorithms in large data systems. Data encryption massive datasets straightforwardly in the cloud presents a higher threat of private key administration and a substantial computational burden because of the quantity, magnitude, and velocity of modifications. The entire database becomes vulnerable to corruption and theft the moment the cryptographic keys is made public.

The importance of cloud-integrated Internet of Things (IoT) solutions in many fields (business, private industry, home devices, etc.) has led to their increased popularity among academics. Employing compact cryptographic techniques and authentication protocols, this study presents a secured data IoT ecosystem. In the suggested strategy, IoT gadgets are classified as either "responsive" or "non-sensitive." The authors advocate for a combination of public versus private clouds, which they call a "mixed cloud." Utilizing RC6 as well as Fiestel, they encode the, like conventional financial, manufacturing, retail online purchases, public services etc.

LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thought- ful analysis of many authors work as follows.

According to Abudul Wahid Khan [4], they want to provide software provider companies with a safe means of using big information in the cloud. In addition to identifying and discussing the aforementioned security issues, the research will concentrate on validating these concerns and identifying techniques for address- ing them via practical investigation. Additionally, we aim to perform a specific example in the appropriate software vendor's organization, similar to the Model of Capability Maturity Conceptual framework, to determine each vendor orga- nization level of the suggested security architecture, and then support them in leveraging big cloud-based data.

The concerns about security and confidentiality surrounding large data plat- forms are significant, as Hanan E. Alhazmi [5] explains. bigAlso, the majority of the currently available technologies are dependent on external parties, which raises serious concerns about data security. In this paper, they present a large data security system that uses blockchain infrastructure and data segmentation to protect sensitive information. The architecture provides a safe space for exchanging, storing, and sending large amounts of data. Blockchain is in charge of ensuring the safety of all massive data storage, retrieval, accessibil- ity, and financial reporting. When it comes to protecting massive data, previous studies have fallen short.

An innovative paradigm for abnormality identification and prediction in large data systems is presented by Marwa A. Elsayed [6]. Customers of Integrated cloud vendors have access to the PredictDeep architecture as a premium cyberse- curity business intelligence tool from a third party. Logging data is gathered from measurement techniques, and PredictDeep utilizes the benefits of broadcasting data mining, graph business intelligence, and deep learning algorithms to uncover anomalies and suspicious behavior. This method's contributions center on resolv- ing the most pressing issues in creating a reliable machine attempting to learn data acquisition for abnormality prediction and identification in large-scale data sets.

According to Seungwoo Seo [7], the reaction time of large data systems is often slowed down by networking security processes. This study proposes a method to improve network security and throughput in the presence of failures and hostile nodes. In order to meet the desired object time of real-time large data compu- tational requirements, the suggested trustworthy streamed dynamic malfunction strategy efficiently detects rogue nodes from among worker nodes. To ensure that data is processed in the allotted amount of time, the trustworthy broadcasting dynamic failure-compensation strategy determines the appropriate number of con- nections by computing the quantity of clusters that yields the minimum trusted computational burden.

According to Yuanzhao Gao [13], the first step toward fully using attribution in vulnerability management monitoring is to construct a derivation architecture that effectively represents provenance data. Existing attribution approaches, un- fortunately, don't really scale somewhat to huge data environments. This work proposes a big data origination model (BDPM) for information security manage- ment relying on the PROV-DM paradigm, taking into account both the unique features of big data and the necessity for close monitoring.

SCOPE OF THE PROJECT

Scope of Big Data Security for Automatic temper detection is explained below

3.1 Hewlett Packard

Data loss and application disruptions can take a significant toll on organizations of every size. Now, with modern data protection, organizations can keep pace with today's changing hybrid cloud environments and ensure data and applications achieve always-on availability. Modern data protection products from HPE enable you to simplify operations with an intuitive cloud experience, meet demanding SLAs, and neutralize damage from threats — helping customers quickly backup and recover data and maximize the value of their backup data, all while reducing costs.
Reference: https://www.hpe.com/in/en/data-protection-products.html

3.2 G2:

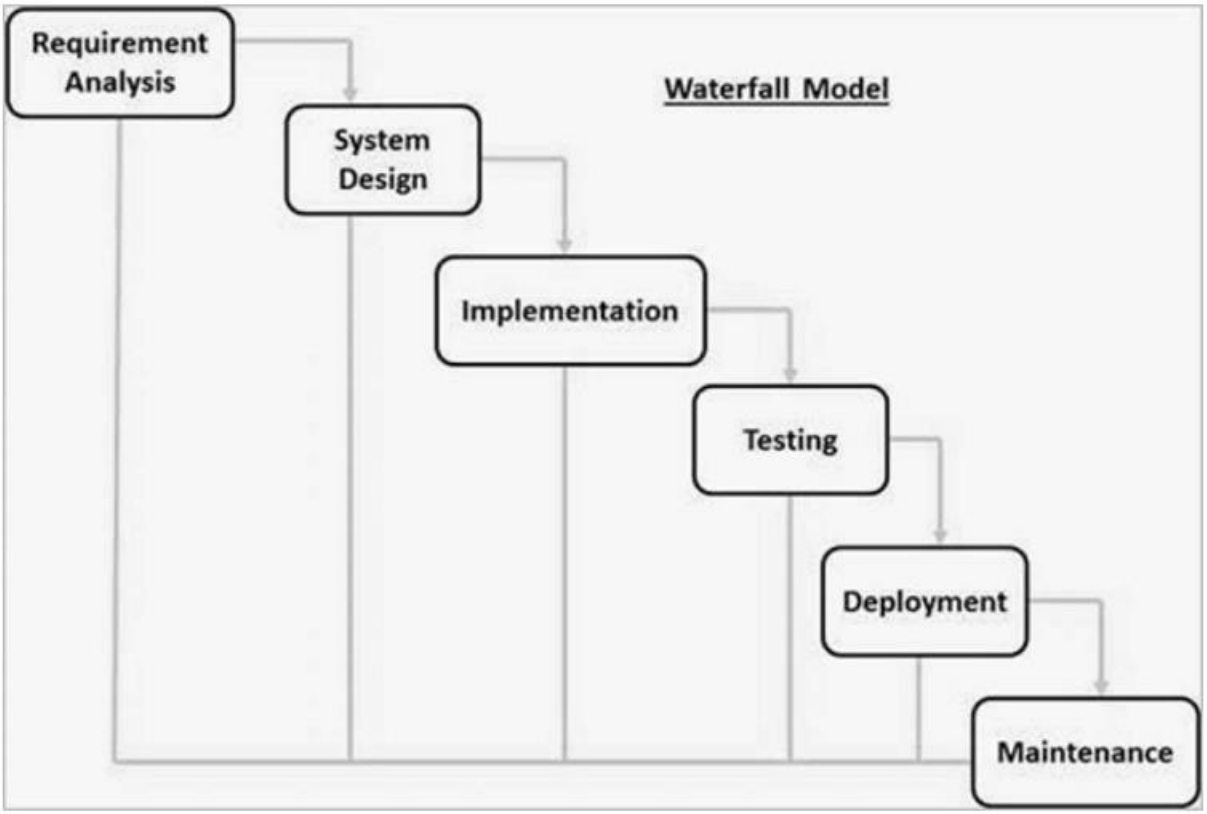
Data security software comes in all shapes and sizes. Tools exist and are designed to secure all types of data, from individual messages to entire databases. Every company, no matter the size, should make data security a core business practice and be doing all they can to ensure data stored in every crevice of their business is protected; any theft to sensitive information can hurt both the business and the client. No sane business owner wants a data breach to be the public's only association with their brand. And no individual wants to provide data to a company known to play fast and loose with their sensitive data. It's important that businesses map out security vulnerabilities and existing security mechanisms to determine where security could be bolstered. Information in every department, from sales to production, should be storing information in a secure fashion and continually updating security measures as new threats emerge. Data security standards may not be required by law in many countries or localities, but information security should remain a priority regardless of their required efforts.
Reference: https://www.g2.com/categories/data-security

3.3 HashiCorp Vault

Data security software comes in all shapes and sizes. Tools exist and are designed to secure all types of data, from individual messages to entire databases. Every company, no matter the size, should make data security a core business practice and be doing all they can to ensure data stored in every crevice of their business is protected; any theft to sensitive information can hurt both the business and the client. No sane business owner wants a data breach to be the public's only association with their brand. And no individual wants to provide data to a company known to play

METHODOLOGY

The methodology for Big Data Security for Automatic temper detection system is developed under waterfall model architecture as shown in the below figure 1.



The sequence phases in water fall model according to our project are mentioned below.

1 Requirement Analysis – Here requirement analysis are done based on following points

- ✓ Base paper for Big Data Security for Automatic temper detection system

2 System Design: The System of Big Data Security for Automatic temper detection system is designed by using the following hardware and software

1 Minimum Hardware Specification:

- CPU : Core i5
- RAM : 8 GB
- HDD : 500 GB

2 Software Specification:

- Coding Language : Java
- Development Kit : JDK
- Front End : Java Swing
- Development IDE : Netbeans 8.2
- Database : MongoDB ,MySQL
- External API : MongoDB API

3 Implementation:

Proposed system is designed by using the following modules

Module A: Data Classification

- Data List
- Labelling
- Unique Data Formation
- Classified List

Module B: Parallel Computation

- Classified Cluster
- Thread Creation
- Data Loading to the thread
- Thread Execution

Module C: Temper Detection

- Vector Hashing
- Hash Mapping
- Avalanche Effect
- Temper Detection

Module D: Forensic Analysis

- Database Lo Analysis
- Temper Details Evaluation
- Data Restoring
- Report Generation

5 Deployment of the system:

The developed software is deployed in the laptop of above mentioned configuration with the help of the mentioned software.

6 Maintenance of the system:

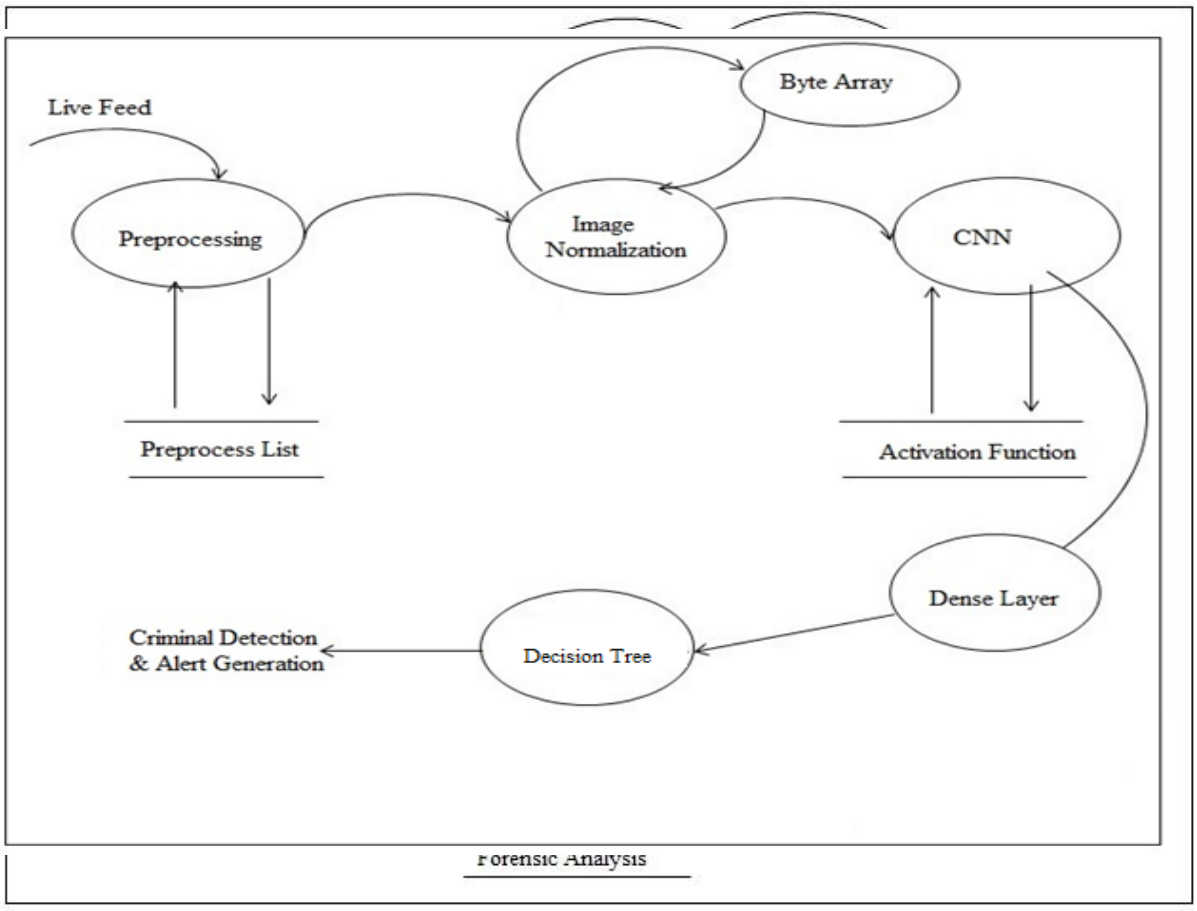
As this software is tested for the quick recovery, so maintenance of the system is not a challenging task. This is because the tools and the software used are open source, so there is no question of licensing the required software.

DETAILS OF DESIGN, WORKING AND PROCESSES

1 DETAILS OF DESIGN

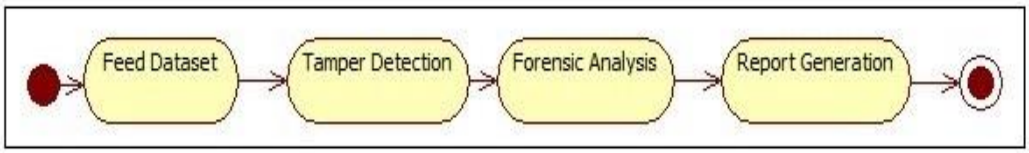
1 Data Flow Diagram

DFD level 2



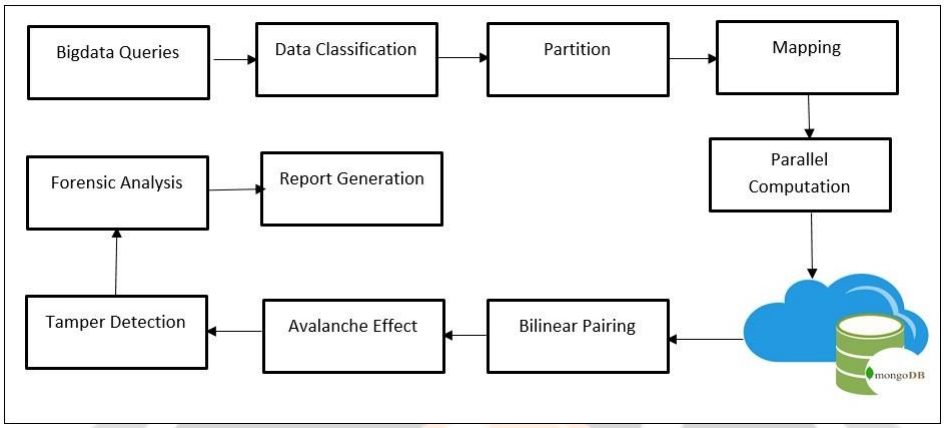
The DFD 2 diagram provides even more details wherein the bigdata queries provided to data classification which further implements the partition. The classified data provided to mapping, which performs the map to tread after which it implements the parallel computation, which is utilize to perform Mongo DB which further do the bilinear paring with Avalanche effect. The temper detection is Performed with the forensic analysis which results report generation.

Activity Diagram



The activity diagram lists the various activities that are performed in the proposed methodology, the start state is initiated and then feed dataset , temper detection , forensic analysis which result report generation and shop activity is done which leads to the stop state.

WORKING AND PROCESSES



Step 1: Query Input and Classification – At this stage of the process, the big data queries that must be processed and run on the database serve as an input that is considered during the process. Since there are many queries, they have to be appropriately categorized into groupings in order to facilitate smoother imple- mentation and execution. At this stage, the input queries are properly organized into groups that are determined by the implementation objective that they are aiming for.

Step 2: Query Partitioning and Mapping – At this stage of the technique, the categorized big data queries that were obtained from the preceding stage are considered to be inputs. These categorized questions are then used to execute partitioning of the queries, which would be done to avoid any conflicts in the queries that might contribute to an error. This procedure is performed in order to complete the partitioning successfully. The appropriate grouping of additional queries that are subsequently offered for performance in an appropriate method is made possible by the functional partitioning.

Step 3: Bilinear Pairing – It is essential that the large amounts of data that are transferred to the MongoDB database be adequately protected from any form of alterations or manipulation. Data that has been tampered with may give rise to security concerns and may result in the disclosure of confidential and personal identifying information data. A hashing technique is used to the data that has been placed on the database in order to get the hash key of the content that has been saved on the database. This hash key is computed and saved at certain durations in the format of a pair-based structure that is referred to as the Bilinear Pairs.

Step 4: Tamper Detection and Forensic Analysis – This is the final stage in the procedure that is carried out in order to guarantee that the authenticity of the large amounts of data that are being saved to the database is preserved. The bilinear pairings that were acquired in the phase before this one are put into practice in order to facilitate an adequate assessment of the system's integrity, which is carried out in this stage. Every time a bilinear pair is formed, it is correlated with itself to look for any variations in the hash string. If these discrepancies are found, it is an indication that the database has been tampered with.

When the tampering has indeed been discovered, the user is informed, and a forensic investigation is carried out to determine the precise amount of time when the tampering was carried out. This is made possible by the cycles of the bilinear pairing that is employed in the process of determining the time frame. A comprehensive summary of the results as well as all of the pertinent data is compiled into the manner of a security report, which is then made available to the user for review.

CONCLUSION

The presented approach for the purpose of achieving the increased security and efficiency of the Big Data through the use of Parallel Computation and Forensic analysis has been elaborated in detail in this research article. The approach initi- ates with the big data queries being fed into the system by the user. These queries are then subjected to data classification which results in the classified data that is then partitioned. The classified and partitioned data is then utilized for the mapping as it will be utilized for the purpose of achieving parallel computation on the MongoDB. The data stored on the MongoDB database is then utilized for bilinear pairing. These bilinear pairs are then constantly evaluated and compared to detect any presence of the avalanche effect. Once the avalanche effect is identi- fied, it means that there has been some tampering that is done to the data. The tampering detection is performed and the forensic analysis is done. This forensic analysis allows for the realization of the report generation. The approach has been quantified through the use of an intensive experimentation that has resulted in highly satisfactory results.

Future Scope

For the future research this project can be enhanced to work in big data in cloud architecture with huge size of the data.

REFERNCES AND BIBLIOGRAPHY

[1] S. Atiewi et al., ”Scalable and Secure Big Data IoT System Based on Multi- factor Authentication and Lightweight Cryptography,” in IEEE Access, vol. 8, pp. 113498-113511, 2020, doi: 10.1109/ACCESS.2020.3002815.

[2] M. Binjubeir, A. A. Ahmed, M. A. B. Ismail, A. S. Sadiq and M. Khurram Khan, ”Comprehensive Survey on Big Data Privacy Protection,” in IEEE Access, vol. 8, pp. 20067-20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[3] F. U¨ nal, A. Almalaq, S. Ekici and P. Glauner, ”Big Data-Driven Detection of False Data Injection Attacks in Smart Meters,” in IEEE Access, vol. 9, pp. 144313-144326, 2021, doi: 10.1109/ACCESS.2021.3122009.

[4] A. W. Khan et al., ”Analyzing and Evaluating Critical Challenges and Prac- tices for Software Vendor Organizations to Secure Big Data on Cloud Com- puting: An AHP-Based Systematic Approach,” in IEEE Access, vol. 9, pp. 107309-107332, 2021, doi: 10.1109/ACCESS.2021.3100287.

[5] H. E. Alhazmi, F. E. Eassa and S. M. Sandokji, ”Towards Big Data Se- curity Framework by Leveraging Fragmentation and Blockchain Technol- ogy,” in IEEE Access, vol. 10, pp. 10768-10782, 2022, doi: 10.1109/AC- CESS.2022.3144632.

[6] M. A. Elsayed and M. Zulkernine, ”PredictDeep: Security Analytics as a Service for Anomaly Detection and Prediction,” in IEEE Access, vol. 8, pp. 45184-45197, 2020, doi: 10.1109/ACCESS.2020.2977325.

[7] S. Seo and J. -M. Chung, ”Adaptive Trust Management and Data Process Time Optimization for Real-Time Spark Big Data Systems,” in IEEE Access, vol. 9, pp. 156372-156379, 2021, doi: 10.1109/ACCESS.2021.3129885.

[8] Y. Liang, D. Quan, F. Wang, X. Jia, M. Li and T. Li, ”Financial Big Data Analysis and Early Warning Platform: A Case Study,” in IEEE Access, vol. 8, pp. 36515-36526, 2020, doi: 10.1109/ACCESS.2020.2969039

[9] M. Li, T. Li, D. Quan and W. Li, ”Economic System Simulation With Big Data Analytics Approach,” in IEEE Access, vol. 8, pp. 35572-35582, 2020, doi: 10.1109/ACCESS.2020.2969053.

[10] C. Bakir, ”New Blockchain Based Special Keys Security Model With Path Compression Algorithm for Big Data,” in IEEE Access, vol. 10, pp. 94738- 94753, 2022, doi: 10.1109/ACCESS.2022.3204289.