

1. Introduction

1.1. Blockchain- Blockchain is a technology that utilizes distributed systems to store data. Each block in a blockchain contains its hash, the hash of the previous block, data and the functions that it uses and stores. The hash of a block is calculated when a new block is created. Through this way, the blocks together form a chain starting from the first block.

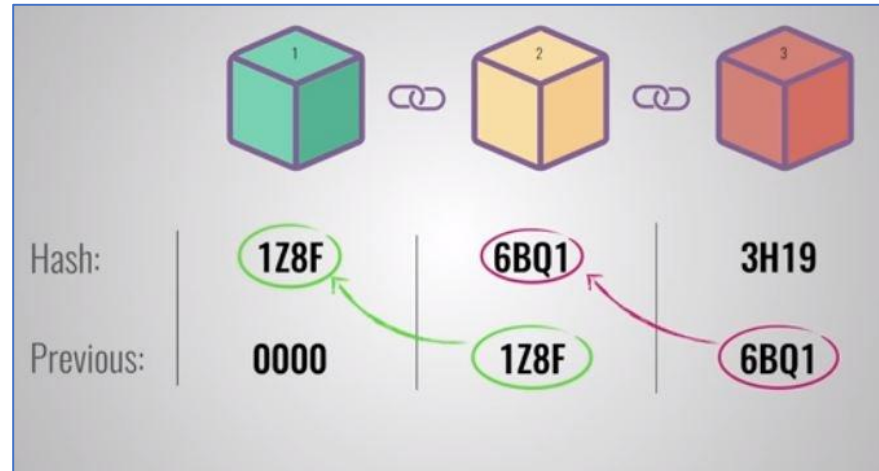


Figure 1: Formation of a blockchain by the links through hash values. [3]

Figure 1 explains the formation of a blockchain. Each block points to its previous blocks and so on.

It is also a distributed ledger which means that the entire blockchain is not in the control of a single entity but every user has its copy. To add a new block to the blockchain, we will have to compute the hash of our block, then we will have to tell about the change to everyone in the blockchain system after giving them proof of elapsed time. After this, every system using the chain update their blockchain and we have a valid system again. This feature makes blockchain very much secure. To change a single block, we will have to take control of more than 50% of the users' systems as we will have to change the blockchain of the more than 50% systems to achieve a majority. We will also have to change the hash of the following blocks after our block in which we intend to make a change. This is because once our block is changed, its hash also changes because of the functions and the programs stored in the block and once the hash changes, the following block no longer points to our intended block, even if we change its hash also, the next one doesn't and the process goes on to the very last block in the chain. This will require a lot of computing power and is nearly impossible to calculate on a traditional computer system. Also, the proof of work makes the creation of new blocks slower to protect this tampering of the blockchain. So, this makes the tampering of the blockchain a long, expensive and tiresome process. This makes blockchain an extremely secure technology.

1.2. Blockchain in Academic Records Systems

It can be used in academics to maintain the records of the qualification of the candidates issued by the institution. These records can then be accessed by the various entities like the recruiters and the candidates when required. They can be verified and be made reliable by using a private blockchain and a combination of username and password. This login system will help to ensure that each entity can access only that amount of data that they are allowed to do so. Also, the system of token generation helps to ensure that the data is exposed to the entities for a fixed time frame only and its privacy is maintained.

2. Background Study

To implement the system, we first read some research papers and noted the merits and the demerits of the current blockchain-based systems in development. The summarizations of the research papers are as follows:

2.1. Blockchain based Academic Certificate Authentication System Overview [1]

Authors: University of Birmingham

The project described aims to resolve the problem of counterfeit academic systems used by various people. It uses blockchain to store that data and specifies functions to maintain the correctness and the reliability of the data being stored in the database.

It has been based on the Blockcerts, which is a project of the Massachusetts Institute of Technology Media Lab. The project provides a scheme of multiple signatures to provide the authentication of the data and the certificates stored in the chain. It uses the hash of the file values to check whether a given certificate is valid or not. It also provides functions to revoke a certificate when the need arises to do so.

The major components of the system are: verification application including federated identity, issuing application involving multi-signature and BTC-address based revocation, Blockchain and local Database adopted by MongoDB.

Figure 2 shows the implementation of the system:

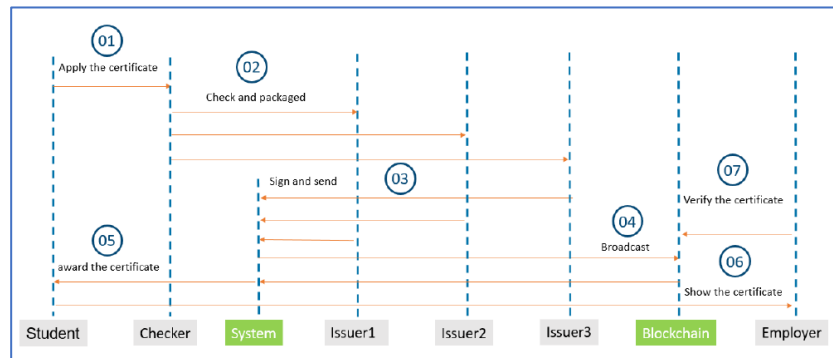


Figure 2: Workflow between different modules of the system [1]

2.2. A Blockchain-Based PKI Management Framework [2]

Authors: Alexander Yakubov, Wazen M. Shbair, Anders Wallbom, David Sanda, Radu State

Blockchain-based PKI (Public Key Infrastructure) inspired many studies proposing blockchain technology to build secure PKI systems. Their main argument is that blockchain-based solution can merge the benefits of the Log-based PKIs and the WoT approaches, and solve some of the problems with conventional PKI system.

The classical PKI systems are CA-based. CAs (Certificate Authorities) issue a signed certificate. For instance, when a user logs into Twitter through a web browser, first the web browser validates the claimed certificate which holds Twitter's public key by checking the CA of the given certificate. Therefore, for a certificate to be trusted, it must have been issued by a Root-CA that exists in the trusted store of the user browser or device, or by sub-CA that has been trusted with Root-CA signature.

In the context of PKI, the blockchain technology provides valuable security features such as certificate revocation, elimination of central points-of-failure and a reliable transaction record. For instance, with the fast certificate revocation, blockchain-based PKIs can instantaneously isolate the infected CA and the corresponding certificates without the need to wait until the next update of Certificate Revocation Lists (CRLs). Also, blockchain-based PKI, as a public append-only log, hence provides the certificate transparency.

Blockchain-based PKI framework supports the revocation of the certificate, which is a real issue in the traditional PKI systems. Moreover, as it is impossible to delete information from the blockchain, only a parent CA can mark a certificate issued by him as revoked. Thus, any misbehaviour of a CA regarding certificate revocation will be also traced and noticed by all other entities.

Advantages of Blockchain-based PKI

A blockchain-based PKI has the following advantages over a traditional PKI:

- The validation of a certificate and its CA certificate chain is simple and fast.
- Blockchain-based PKI solves a longstanding problem of traditional PKIs by not requiring the use of a service that issues certificate revocation lists (CRLs) due to blockchain synchronization between network's nodes where any modification to the state of a certificate will be instantly notified to all nodes.

3. Requirement Analysis

3.1. Introduction

3.1.1. Description: The project aims to implement a distributed database system using the concept of private blockchain implemented through an authorized user login system. It works on P2P architecture and each node in the network can directly interact with each other.

3.1.2. Environmental Characteristics

3.1.2.1. Hardware: The system will be able to run and access the records on normal PCs and servers. The suggested PC hardware requirements of such a system are as follows:

1. RAM: 4GB or higher
2. Secondary memory: 512 GB or higher
3. Processor: Any processor of Intel 3rd generation or higher with 4 or more cores.

3.1.2.3. Software Requirements: The system will require a Windows Operating System with a version of 7 or higher. The system will run using Python as the primary language and using “Jupyter Notebook” Software. This software will be required as the program uses “Pandas” and “Numpy” libraries.

3.1.2.2. Actors: The stakeholders involved in using and maintaining the system are the institutions and regulatory authorities that are participating and updating and adding new records. The other stakeholders are the candidates and the recruiters.

3.2. Functional Requirements

- 3.2.1** Maintain academic and professional records.
- 3.2.2** Provides easy traceability of the data by searching for hash values within a chain.
- 3.2.3** Valid users are provided with a username and password for login.
- 3.2.4** Tokenization systems help to ensure data is not exposed for too long.

3.3. Non-Functional Requirements

- 3.3.1** The timestamp should be maintained which contains date and time at which the block was created.
- 3.3.2** Maintain hash of the latest block so that a new block can be added to the chain using the hash value of the previous block.
- 3.3.3** Implement the above project using a private blockchain.
- 3.3.4** Allow only authenticated users to update or add record.
- 3.3.5** Data or a block can only be added in the chain if and only if it is verified by the majority of the existing blocks.

3.4. Use Case Diagram

Figure 3 represents the use case diagram for our system.

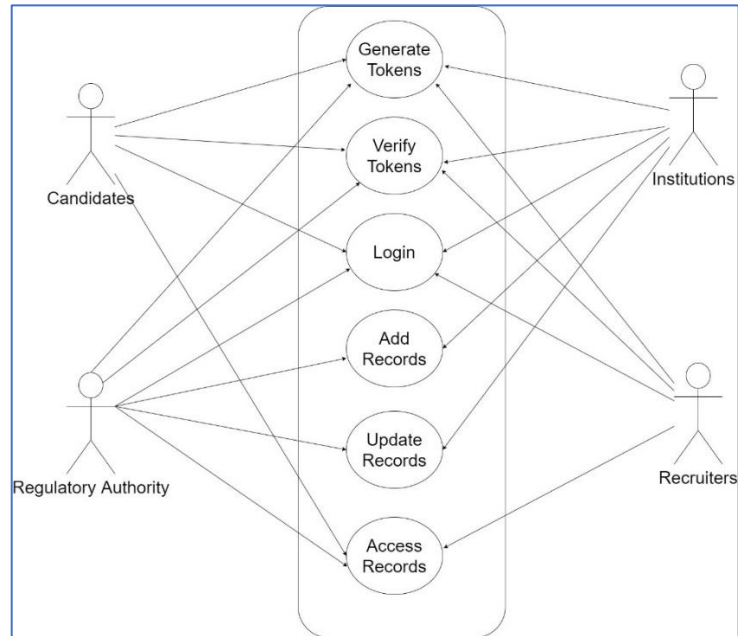


Figure 3: Use Case Diagram

4. Detailed Design

Figure 4 shows the structure of the system

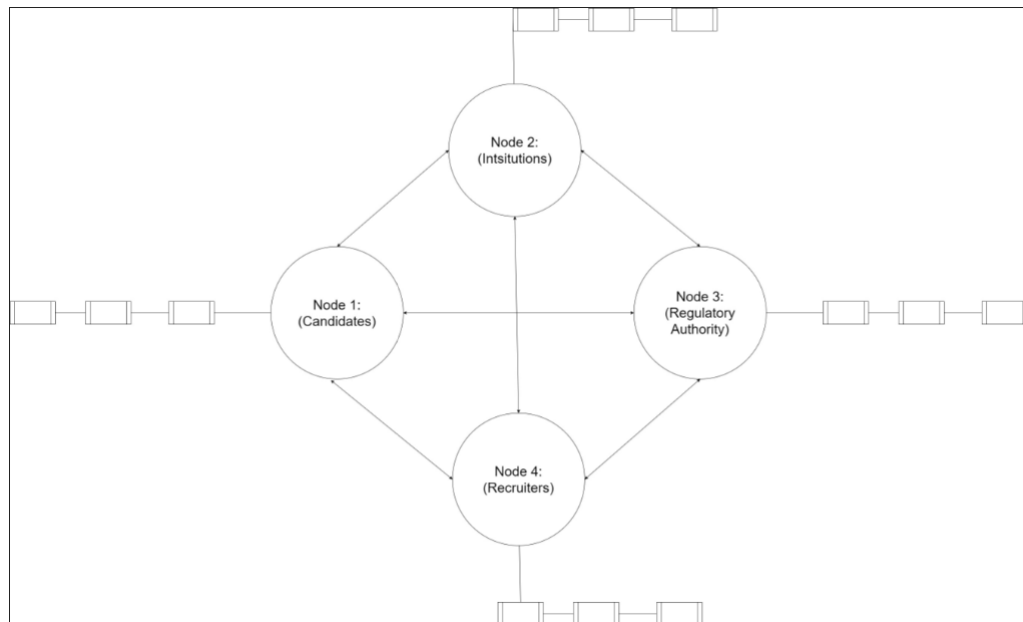


Figure 4: Representation of the nodes in the network

Each node represents an entity which is depicted by the circle and contains its blockchain copy.

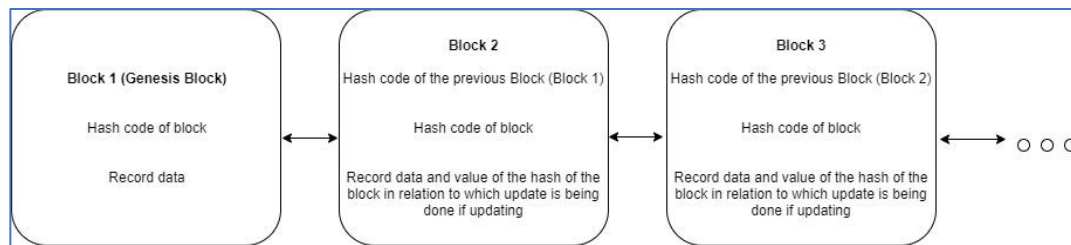


Figure 5: Diagram representing the blockchain structure in the form of blocks

The first block (as shown in figure 5) is called the genesis block and it contains only its hash and data and it will not have any update data as it is the first block in the chain. After that, the second block contains its hash, the hash of its previous block, i.e., the genesis block and data. In the data field, the following attributes are stored:

1. Update data and the hash of the block about which the update is being done if any.
2. The candidate qualification records. This will contain the following fields:
 - 2.1. The name, age, address and contact details of the candidate.
 - 2.2. The academic qualifications of the candidate with the complete details of the degrees and the institutions that have issued them.
 - 2.3. The work experience of the candidate along with the certification and testimonials issued by the organizations that the candidate has worked in previously.
3. A token generation function for ensuring the privacy of the data that is being accessed. These will contain the following features:
 - 3.1. The token generation function will generate a token that will consist of a numerical value when the request for data is made. The function will be so written that the token generated by both the requesting node and the function of the chain will be the same. These will then be matched. If both the tokens are the same, then the chain will allow the data to be accessed. Also, this will be for a limited time frame. This will help to provide an additional layer of privacy for the system other than the login system.

4.1. Features of the system and the problems that will be solved

The main features of the project and the major problems that this project will be able to solve are:

4.1.1. Traceability

The entire records of the database can be searched through the database system easily. Each update is added as a separate block. Also, it means that if we have to trace where the change or addition of a block originated from, we can easily do so as the updating will contain the hash of the block where the change was made. This means that tracing changes to a record of a student in case of correction by an institute or regulatory authority will be

far easier and more efficient as you will not have to search the entire database to check for the change.

4.1.2. Security

The key advantages of a blockchain-based record-keeping system are as follows:

- The blockchain itself provides a very strong layer of security for the entire database. Every new node being entered has to be verified and updated in all the copies of the blockchain. Also, only certain users authorized by a login system can add new nodes to the database. This ensures that no outsider can access the information.
- To change any particular block of information in the entire blockchain, you need to have control of more than 50% of the entire chain. Only then can you prove verification of the entire chain with the changed block. A single change of hash at any particular block will make the entire blockchain invalid.
- Further, the system of tokenization ensures that the data is exposed for a very limited time frame and does not remain in the control of the requesting entity for long. This helps to decrease the network congestion and it will also help in reducing the chances of the data leaks.

4.1.3. Credibility

One of the main issues with the traditional database systems is that of credibility. Since these records are more prone to hacking attacks and are not considered safe, the various entities involved in the system do not trust them. For example, if a company wants to recruit candidates, then it is very possible that the company might not trust the database stored on the central servers as the data on them can be modified by various other players. So, the company decides to collect its information. This leads to additional costs for the company. However, a distributed database system such as the one provided by the blockchain will not be easy to modify. There will be multiple copies of the chain stored on many participating nodes. These all will act as points of verification for the entire record. Any entity that needs to verify the data can use any of them or all for verification. Hence, the system reduces the costs for everyone involved in the supply chain.

4.1.4. Transaction Processing

The transaction processing, i.e., updating and adding new records to the database will be far more efficient. We are including a new block to the chain with each update and this means that we don't have to overwrite any existing data values to change the previous values. This results in less time being wasted to search and then add another node. Also, if we don't need to delete any previously stored data if we have to change it.

4.1.5.Privacy

The given system also helps in improving the privacy of the data being stored as it is available for a very limited time frame only and for multiple data requests, multiple tokens have to be generated. The number of tokens that can be generated is also limited for a time interval. This helps to ensure that none of the nodes can try to extract huge amounts of data from the database for malicious use.

5. Implementation

Each block will contain three main sections:

1. **Timestamp:** This will contain the date and the time at which the block was created. This helps to identify the relative position of the particular block in the blockchain. It will also help in tracing the blockchain whenever a search needs to be done and there is no data available for the search other than some basic fields. The timestamp can be given as input and then, an appropriate search algorithm can be used. For example, if we are searching for a block with a timestamp as 29-03-2019 05:06:23 and the middle block of the chain is with the timestamp of 05-06-2018 12:00:00, then we don't have to continue the search after the middle block. This will reduce the time required for searching.
2. **Hash of the Previous Block:** The hash of the previous block is crucial and it denotes the previous block in the chain. This is decided after the new block to be created has been verified and accepted by a majority of the participating nodes in the network. After the verification flag is set for the incoming block, it is communicated to all the nodes so that they all can add to their copies of the blockchain.
3. **Hash of the Block and the data:** The rest of the block comprises of the data that is to be stored in the database and the hash of the block. The hash of the block is calculated at the time of the block creation. The data contains the records of the candidate. Also, if the block was created about the update to the previous block, then the data will also contain a field as "Update: ... time" where the three dots denote the hash value of the block about which the update was added and the timestamp of that particular block.
4. **Tokenization system and the authorization functions associated with it:** A function to generate tokens is run on each node. It is run in two phases after a request for data access is made. In the first phase, the function runs on the node side and the token is generated. After this using a consensus mechanism, the token is generated on the chain software also. If the tokens match, then both the sides will agree to share the data. This is also allowed for a limited time only and once that time expires. The data can't be accessed after that.

This ensures the privacy of the data and ensures that the data is not exposed to the participating entity for more than a decided time to avoid data breaches. This makes the system extremely secure and reliable.

Figure 6 describes the whole process:

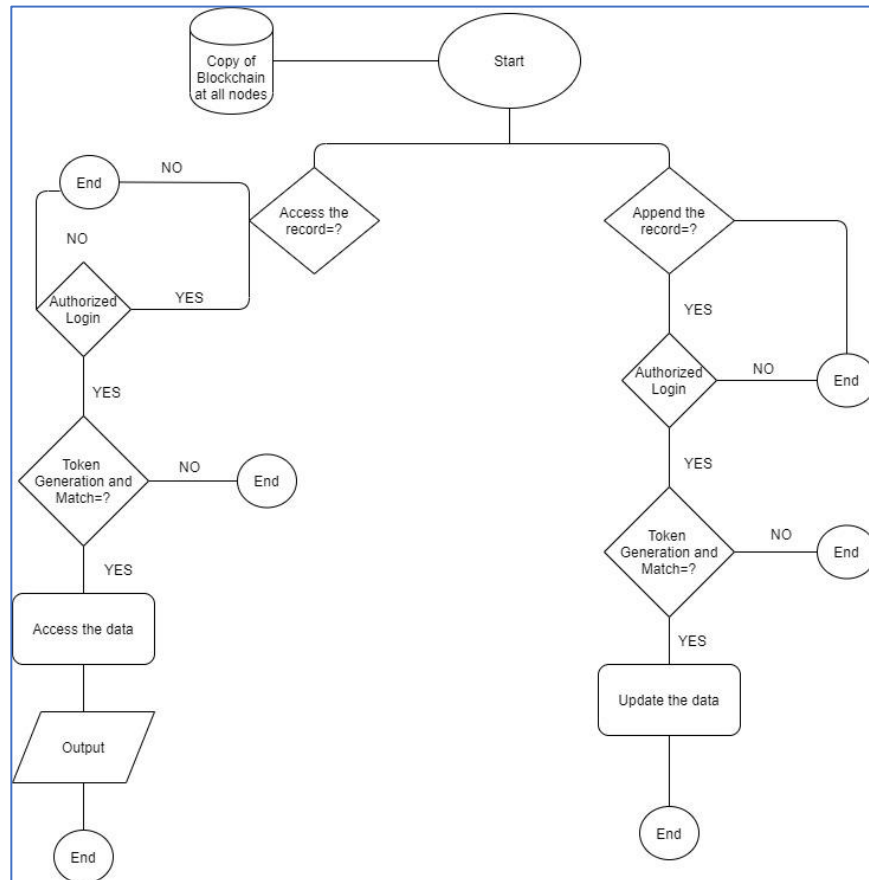


Figure 6: Flowchart of the system

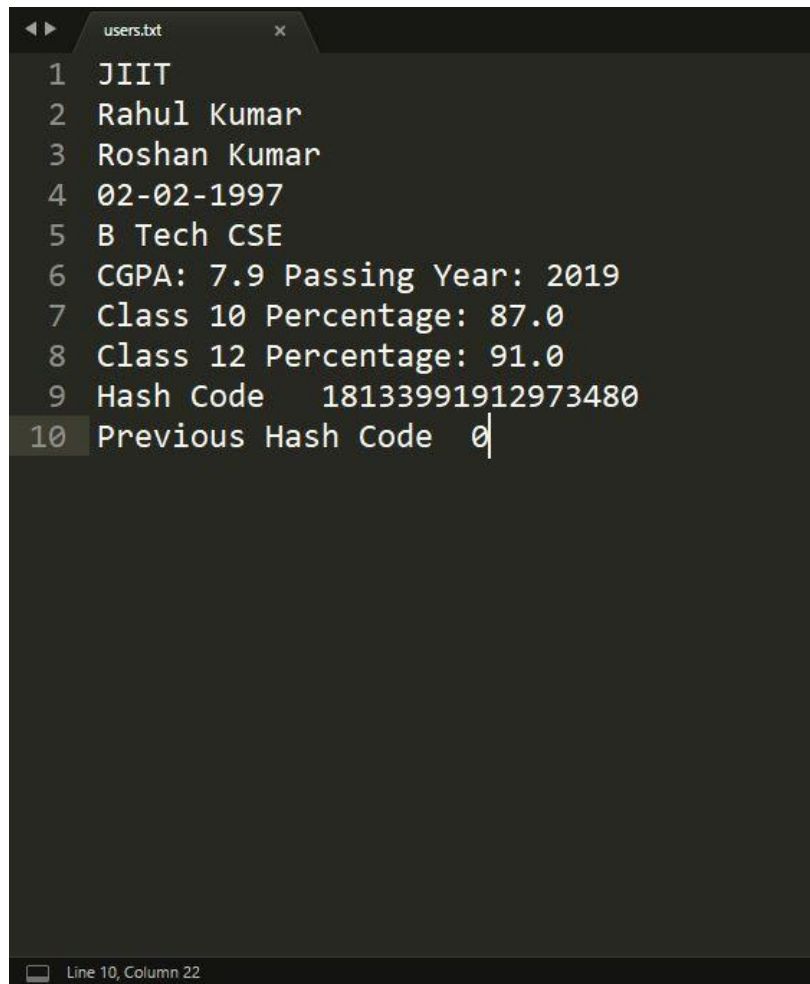
All the functionalities mentioned above have been implemented using the Python 3.6 language and have been tested using the “jupyter Notebook” software on Windows 10. However, the programs of the system will be able to execute correctly on almost all the common Operating Systems provided they have an installation of Python and Jupyter.

6. Experimental Results and Analysis

The outputs of the various sections of the code are:

```
The data in the user file is as follows:
[['institution' 'abhishek' 'bhati']
 ['recruiter' 'shivam' 'goel']
 ['candidate' 'kushagra' 'aggarwal']]
The shape is as follows:
(3, 3)
Enter your choice from the options given below:
1. Enter new records
2. Access the records
Enter your choice1
Enter your username:abhishek
Enter your passwordbhati
User accepted.
Enter the new records
Enter the name of the institution which is issuing the record:JIIT
Enter the name of the candidate:Rahul Kumar
Enter the father's name of the candidate:Roshan Kumar
Enter the Date of Birth of the Candidate:02-02-1997
Enter the course Details for which the certificate is being issued:B Tech CSE
Enter the passing details of this particular course:CGPA: 7.9 Passing Year: 2019
Enter the no of old records for the candidate that you want to enter:2
Enter the details of the record:Class 10 Percentage: 87.0
Enter the details of the record:Class 12 Percentage: 91.0
```

Figure 7: Output: Addition of records to the Chain after verification



```
users.txt
1 JIIT
2 Rahul Kumar
3 Roshan Kumar
4 02-02-1997
5 B Tech CSE
6 CGPA: 7.9 Passing Year: 2019
7 Class 10 Percentage: 87.0
8 Class 12 Percentage: 91.0
9 Hash Code 18133991912973480
10 Previous Hash Code 0
```

Figure 8: Output: Input Data written in the Chain File

```

The data in the user file is as follows:
[['institution' 'abhishek' 'bhati']
 ['recruiter' 'shivam' 'goel']
 ['candidate' 'kushagna' 'aggarwal']]
The shape is as follows:
(3, 3)
Enter your choice from the options given below:
1. Enter new records
2. Access the records
Enter your choice2
Enter your username:shivam
Enter your password:goel
User Accepted.
Accessing the records for 10 seconds
JIIT
Rahul Kumar
Roshan Kumar
02-02-1997
B Tech CSE
CGPA: 7.9 Passing Year: 2019
Class 10 Percentage: 87.0
Class 12 Percentage: 91.0
Hash Code      18133991912973480
Previous Hash Code      0
The access time has elapsed.

```

Figure 9: Output: Implementation of Tokenization in the access function

After executing various test cases on the system, we have made the following observations for the system:

- The system properly uses a tokenization system to limit the access time of the stored data to provide the immutability of the records in the system and make the data safer.
- The system has implemented all the features as detailed in the Design part of this report. All the modules are working cohesively with each other to generate the desired output.
- The system can transfer all the files in between the various nodes of the system according to the desired specifications.
- The hash functions can provide a unique hash value using the complete data present in the files generated by all the nodes after taking the user input. This has made tracing the previous records and updating them faster and easier thereby increasing the performance of the system.
- The errors and prevention of unlawful access to the data stored at the nodes are working by using a combination of both the login system and the tokenization system.

7. Conclusion of the Report and Future Scope

- The given system has enabled the user to store the data on a blockchain in a secure manner with controlled access to each verified user and this has enhanced the security of the data access and manipulation in comparison to the traditional ways of storing the data on the servers.
- The system is also able to provide a secure way of authentication and this can decrease the amount of data problems as mentioned previously in this report, thus requiring fewer resources and maintenance cost compared to the systems that are available in the market right now for storing the data of the candidates.
- The system will be able to address new use cases as and when they arise as it will be very easy to issue updates on a distributed system as compared to a centralized system for data storage. The update feature will be very useful to keep the hash functions more and more secure by changing them whenever the need arises. Hence, the system can be further enhanced in the future with the help of more functionalities added.
- More participating entities in the system can be added in the future to the project. This can include government authorities which can overlook all the procedures. This will further reduce the workloads on their servers too and will help in enhancing the usability of the system.
- Rather than using the file system for storing all the data, a much more robust database system can be used in the future implementation of the project and this will help to make the system more secure in the future.
- The data stored at each node can be encrypted with the right algorithms balancing between the processing power and complexity to add an extra layer of security for each node.

8. References

[1] Blockchain based Academic Certificate Authentication System Overview

Li, R., & Wu, Y. (2018). Blockchain based Academic Certificate Authentication System Overview.

[2] A Blockchain-Based PKI Management Framework

Yakubov, A., Shbair, W., Wallbom, A., & Sanda, D. (2018). A blockchain-based pki management framework. In *The First IEEE/IFIP International Workshop on Managing and Managed by Blockchain (Man2Block) colocated with IEEE/IFIP NOMS 2018, Taipei, Taiwan 23-27 April 2018*.

[3] Online:

<https://sites.google.com/site/debasish22blog/blockchain-basics>