A Framework for Securing Data Warehouse Using Hybrid Approach

Queen P. Kalio & Nuka D. Nwiabu
Department of Computer Science,
Rivers State University,
Nkpolu – Oroworukwo, Port Harcourt,
Nigeria
angelisticqueen95@gmail.com

Abstract
Data warehouse stores enormous quantity of data for many years. Hence, the security and safety of this data/ information is very important and it is critical in the maintainability and regularity of stored data. Stored data has encountered a variety of technological changes, which has provoked some strategic improvements in the security policy of the system. Threats to the security of data warehouse by intruders are detrimental to data owners. The insecurity of the data warehouse has caused organizations huge cost especially in the event of data theft. This study presents a framework for securing data ware house using a hybrid approach and was achieved by identifying security factors of the data warehouse creating hybrid security model with 3 level authentications. The system is web based that uses email and password, Token, CAPTCHA authentication and verification to ensure strict adherence to this process effectively and efficiently.

1. Introduction

Data warehouses (DW) are centralized data storehouses that incorporate data from varieties of transactional, legacy, or visible systems, applications, and fountain. The warehouse simply provides a background or setting distinct among functional systems which is entirely intended for the purpose of analytical-reporting, decision-support, data mining and unplanned queries. The optimized separation facilitates enquiries to be carried out without any contact on the systems that support the business' principal transactions (Dedić & Stanier 2016). The type of data in the repository could define a Data warehouse, and the people that accesses it for use and it is intended for decision support in business activities. This warehouse is separated from daily online transaction processes and the applications that compel the heart of businesses, reducing disagreement for both active businesses and analytical queries (Inmon, 1991). According to (Caserta & Kimball 1998), the warehouse is usually read-only and its data systematized agreeing to transaction requirements than computer processes. The warehouse rank information in accordance to importance to transaction managers and analysts, for instance, products, customers and account. Data warehouses store vast amount of historical information ranging from very distant to very recent transactions. For this purpose, data in the warehouse is frequently condensed or combined making it uncomplicated to access, scan and query. Frequently, disused data are incorporated in a warehouse to provide users with multiple views of information which is being presented in logical and easily understood clustering (Kimball et al, 2013). In general, Data warehouse facilitates business organization in analysing its business data, whereas the operational database that supports the production that a system engaged in the business process and contains working data that are atomic in nature. Essentially, it helps resolve the present-day issue of getting data out of systems promptly and professionally and translating the raw data into useful information. A Data Warehouse is an essential part of an organization and authorizes its users, enabling them to retrieve information

concerning business process as a whole. Devbandu, and Stubblebine, (2000) in their work said, security is a significant requirement in Data Warehouse application, starting from the gathering of requirements through implementation of the system and maintenance. (Inmon, 2002: Yuhanna, 2010).The aim of this study is to develop a framework for securing data warehouse using a hybrid approach of user name and password, token and CAPTCHA.

2 Related Works

Security concerns have been an important subject in the corridor of data warehouse system which remains totally unresolved. This chapter reviews related work from previous studies carried out by other researcher on the data warehouse system. The studies are reviewed below. **Okerenke, (2015)**: provide a hybrid framework for security issues in Data warehouse (which comprises the use of Username/Password and Token generation). The bottleneck in this work was that the system can only use password-based verification where Usernames are often a mixture of the individual's first name and last name, this makes them guessable. If restrictions aren't put in place, people frequently produce weak passwords -- and even passwords that are strong enough can be stolen or forgotten. Password authentication weaknesses can be addressed to some extent with smarter usernames and password rules like the length and conditions for complexity, such as including capital letters and symbols. Nevertheless, password-based authentication and knowledge-based authentication (KBA) are more vulnerable than systems that require multiple independent methods. Similarly, **Sajjad et al, (2014)** said that to provide a hybrid framework for security issues in Data warehouse (which comprises the used of Username/Password and Token generation); the new hybrid technique centres on recognizing the level of users. Identifying of various levels of users initiates the proper categorization of users as they are placed to gain access to data. The access control of the system is made much safe and secured by classifying the users. This division of users also makes sure that the authorized users can only access the data. Unauthorized or immaterial users cannot gain access to data that is not of concern to their functions but still, intruders had access at the back end of the software. Also, **Sreedhar et al, (2011)**: used the Schematic method using Data type Preserving Encryption to enhance DW Security, the first stage of his work was to secure the Data warehouses using the basic data type prevention technique afterwards the enhanced encryption technique is used with DES 56 key, also discussed with the other preventive mechanism for Data warehouses security, the following measures are used to boost Data warehouses security. (Prevention, Detection and Containment, Recovery, Investigation, blinder key methods also used). However, **Colin et al, (2012):** developed a tool for tracking the data lineage and create data flows successfully and efficiently on a heterogeneous environment. They were able to Captcha high-level patterns by clustering trace log entries and discovering chronological associations between clusters. Therefore, the tool is able to accurately discover flow and sequence structures. Similar study on **Somchart et al, (2009)** gave an approach for a structure of collaborative admittance control for OLAP queries spanning over multi-Data warehouse founded on PKI and PMI, RBAC and multi-agent system. They are showed how multiple Data Warehouse security policies are incorporated and handled through the XACML document specification. Finally they provide the implementation of the prototype A-COLD. Microsoft SQL Server White Paper Released on April 2102 gives the comparative study of security features in SQL Server DWs Concepts with MySQL 5 DWs concepts, in this comparison there are 9 security measures are taken out of 9 only three measures are supported by both SQL Server 2012 and MySQL 5, other than three are supported by SQL Server 2012 Only. Similarly, **Juelog LI et al, (2009):** provides a Data warehouse security in field service level at the application of flight service. In using the process of dimensional modelling to develop FSP Data warehouse, putting original data in the different departments through clean and conversion, then stores in Data warehouse. For the DSS of flight operations the data can

be retrieved from the FSS with security. **Hallman et al, (2012):** concluded security issues about Data integrity as, integrity on the whole is the most potential unpredictable part is the success of any database. A-well designed and maintained database can guarantee key area, and referential reliability. Ease of access and network security tools are a very imperative aspects of data management activities. At a core he said the data integrity is maintained by the users with the common knowledge of their domain or system with the working skills of computers. However, **Ricardo et al, (2010):** proposed a data masking solution to enhance data privacy in Data warehouse. He also uses one of the covered fact tables masking key for facilitating false data injection and raising the security strength against attackers. Also, **Raj et al, (2014):** they worked on Data warehouse system which finally suggested the security measure to prevent the sensitive data from malicious attacks. He provides a log based security system architecture to prevent the data from the attackers. David et al, (2014): conclude his comparative work of design approaches giving heed to security perspective; not a hint of the approaches will start off securing a Data warehouse development from requirement level to final implementation. Bernardino et al, (2012)**:** he has concluded that the existing encryption resolutions are not suitable for warehouse technologies. Data warehouses function in a well-determined explicit location with rigid security, performances and scalability requirements and consequently, need specific solutions that are able to survive with these directives. Also, (Carlos et al, 2009): provide a MDA approach that supports a protected design of Data warehouses. This architecture to begin with, only supported a rational pathway towards DBMS, but since majority of Data warehouses are administered over a multidimensional approach, it was recently enhanced with a multidimensional pathway in the direction of OLAP tools. The security rules to the MDA architecture, by adjusting our conceptual met model to stand for these security rules and defining sets of QVT transformations which automatically create multidimensional logical models from conceptual model. In addition, (Thangaraju et al, 2014): discussed the security and performance in ETL with the transformations operations finally conclude the unconnected lookup transformation is the best for security and fast transformation in a well design organization. The idea introduced by Vassiladis, Bouzegeghoub and Quix, (2000), in whose approach covers the full lifespan of the data warehouse, and consent to capturing the interrelationships that exist between different quality factors and helps the interested user to organize them in order to fulfil specific quality goals. Furthermore, they prove how the quality management of the data warehouse can guide the process of data warehouse evolution, by tracking the interrelationships between the components of the data warehouse. Finally, they presented a case study, as a proof of idea for the proposed methodology. The concept introduced by Santoso and Gunadi, (2006), their paper describes a study which explores modelling of the dynamic parts of the data warehouse. This meta model permits data warehouse management, design and evolution based on a high level theoretical perspective, which can be related to the actual structural and physical aspects of the DW scheme. Besides, Meta model is competent in modelling multifaceted conducts, their relationships between data sources and execution details. The paper introduced by Nizar et al, (2010), The aim of this paper is to discover the main critical success factors (CSF) that leaded to an efficient implementation of DW in different organizations, by comparing two organizations namely: First American Corporation (FAC) and Whirlpool to come up with a more general (CSF) to guide other organizations in implementing DW efficiently. The result from this study showed that FAC Corporation had greater returns from data warehousing than Whirlpool. After that and based on them extensive study of these organizations and other related resource according to CSFs, they categorized these (CSF) into five main categories to help other organization in implementing DW efficiently and avoiding data warehouse killers, based on these factors. The paper introduced by Manjunath T.N, Ravindra S Hegadi, (2013), The proposed model evaluates the data quality of decision databases and evaluates the model at different dimensions

like accuracy derivation integrity, consistency, timeliness, completeness, validity, precision and interpretability, on various data sets after migration. The proposed data quality assessment model evaluates the data at different dimensions to give confidence for the end users to rely on their businesses. Author extended to classify various data sets which are suitable for decision making. The results reveal the proposed model is performing an average of 12.8% of improvement in evaluation criteria dimensions with respect to the selected case study.

3. Methodology

To carry out this research and achieve the task, the experimental research method and object – oriented analysis/ recursive design was used. The Experimental Research is a research approach that determines result amid conflicting hypotheses. Experiment is a useful tool for research investigation since it allows researchers some control in influencing variables under tight condition (Blaxter et'al, 2006). And the Object Oriented Analysis/ Recursive Design were used as the system design. This entire system is broken down into subsystems and modules, and the modules are seen as objects. The Object-Oriented Analysis / Recursive Design is basically an iterative process of carrying out object oriented analysis, system modelling pre and post analysis for different systems. In computer science, Recursion helps in solving problems where the solution depends on solutions of smaller instances of the same problem as opposed to iteration (Donald and Oren, 1990). This method followed, helps in successfully designing this system knowing that, these models has the ability to separate different subject matters while building application independent design. And a hybrid method of data collection was adopted for achieving the objective.

4. Design Approach

This specifies what the developed program is supposed to do in terms of its functionalities and performance. These functionalities will be based on the proposed framework for securing data warehouse. Before the ware house can be accessed, one need to go through three (3) tier authentication securities checks/pass. Firstly, the user has to provide legitimate login details for authentication. If the username and password is not legitimate, an illegal access message pops out alerting the user that the credentials provided is “incorrect email and password combination”, then, if the user details are correct with the one in the database, a token (i.e. one time pass code) is forwarded to the customer’s registered email. When the token has been verified and passed, the user will now have to deal with the CAPTCHA security pass to access the warehouse. If the user is not able to go through, the system denies access to the user to access the warehouse.

4.1 Use Case Diagram

This shows the relationship between system’s users during implementation process. It presents different use cases the actor (user) is involved with. It shows pictorially the functions and their interactions. The figure 3.4 shows the communication connecting the system and its execution process.

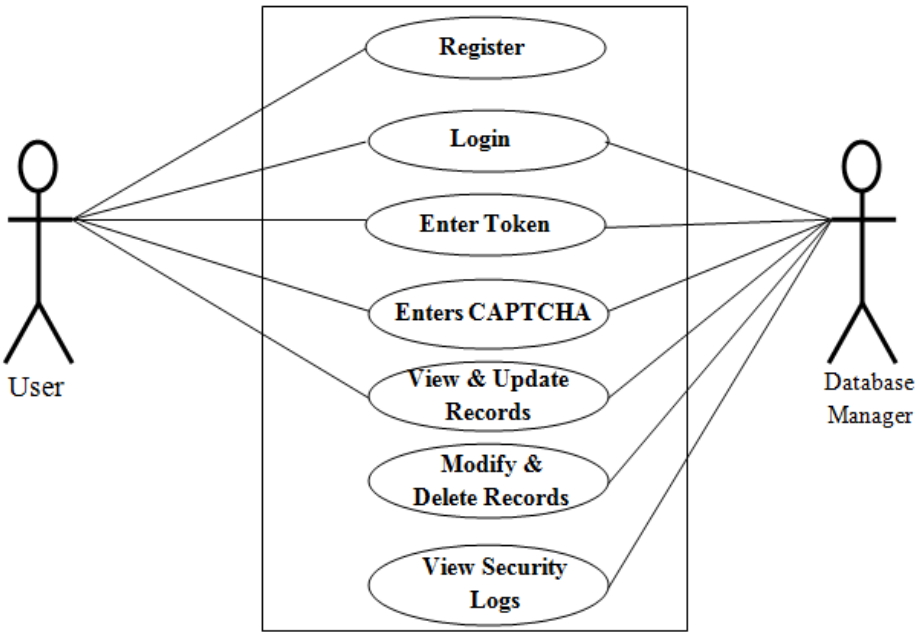

Figure 4.1: Use- case diagram of the proposed system

4.2 Architecture of the proposed system

This work is based on the proposed framework for securing data warehouse. Before the warehouse can be accessed, one need to go through three (3) tier authentication security checks/pass. Firstly, the user has to provide a valid email and password for authentication. If the username (email) and password is not valid, an error message pops out alerting the user that the credentials provided is not correct (i.e. incorrect email and password combination), then, if the username and password is correct with the one in the database, a token (i.e. one time pass code) is sent to the user's registered email. When the token has been verified and passed, the user will now have to deal with the CAPTCHA security pass to be able to access the data warehouse. If the user is not able to go through, the system will not allow the user to access the warehouse. The figure 1 displays the architecture of the system.

4.2.1 Architectural Design of the proposed system

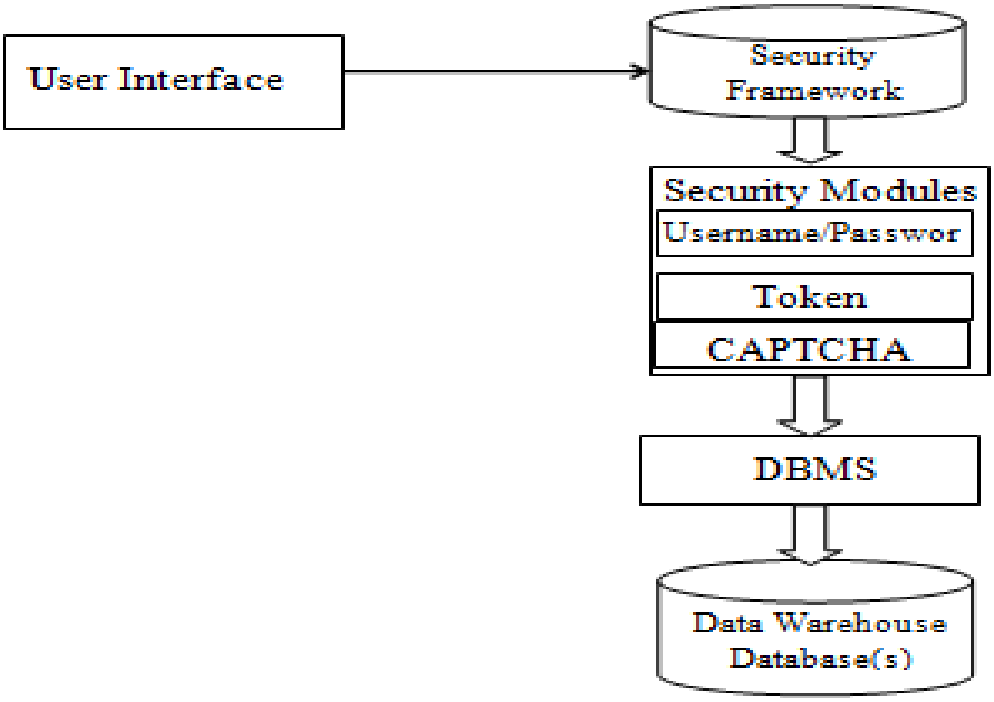

Figure 4.2: Architecture of the proposed system

A. User Interface

This is a user friendly screen through which user can gain access to other parts of the system. This interface lets the user submit login details to gain access into the system.

B. Security Framework

The Security Framework is all the measures put together to protect and secure a database or its warehouse from illegal entry or illegitimate use and malicious attacks. Database enables the identification of each Data Warehouse user through their access details stored in the database and data access policies which are the attributed role(s) and SQL grant privileges. The Security Framework Database stores all the necessary OTP verification code and Captcha for each Data Warehouse database that needs to be masked in a log file. Also, the Security Framework Database stores all the Data Warehouse user behaviour profiles that will be used to assess the incoming user statements. The rule base for the rejected method and the rejected measure computed for each authentication technique is also stored in the access protocol.

C. Security Modules

The Data Warehouse Security level Interface is being used for access control for securing the Data warehouse. The proposed system follows these actions by the sub components to restrict access to the warehouse. These sub components includes: the username (email) and password, token (one time password) and captcha.

a. Username and Password Security Approach

The username and password is the only process that is supplied by the user during registration. After the user registers into the system, information about the user is store such as name, username, chosen password, email address etc. The username and password are requested for during login, when the system has successfully authenticated the user, then the system can proceed to the next level.

b. Token Verification (One Time Password) Security Approach

A single-use password or series of codes used to authenticate a user over an un-trusted communications channel. The one time password system was implemented to add extra layer for the authentication of users. The following are the process in which this technology works:

1. The user first registers on the platform
2. While trying to login, the user will pass the email address and password authentication.
3. In the backend, a code will be generated and sent to the user's registered email address.
4. Users key-in the OTP code.
5. In the backend, once the code is verified, the user will be taken to the next stage.

c. CAPTCHA Security Approach

This is the 3rd and last security authentication process of the proposed system and it's an acronym which stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart. The general logic behind the CAPTCHA figures and letters generation is:

1. It randomly generates a number and character string from strings of 1 – 9, a – z and A – Z.
2. Then inserts the same string in the session and also creates an image and write that string to image.
3. So, when browser reloads, it refreshes session random string and also the CAPTCHA characters.
4. On form submit, it validates the posted value with the session values.

D. Database Management System (DBMS)

DBMS is the system in which user information are stored. The MySQL language is used in the manipulation of the information.

Table 4.1 Database of the proposed system

S/NO	FIELD NAME	DATATYPE	FIELD SIZE
1	matricNo	Int	15
2	First Name	Varchar	20
3	Last Name	Varchar	20
4	Level	Int	3
5	Department	Varchar	50
6	Faculty	Varchar	30
7	phoneNumber	Int	20
8	Address	Varchar	50
9	City	Varchar	12
10	State	Varchar	15
11	Country	Varchar	8
12	Email	Varchar	30
13	DoB	Date	10

E. Data Warehouse

The Data Warehouse Interface is the compartments were all forms of raw data; Meta data or processed data can have access to and manage. When an authorized user scales through all the security protocol and is able to get here, he/ she can perform the task allowed to him/ her.

A frame work for securing data warehouse as shown in figure 2 gives an illustration of two actors. One is a data warehouse 'user' and the other is the 'database manager' that manages the warehouse and checks for security bridges by viewing the security logs, modify and delete records. He also go through the security protocol to access the warehouse and to view and update records. The user follows the security protocol during login to view and update their records.

5. Result and Discussions

5.1 Results

The system was able to perform the required security checks and it record about 100% success as there was no failure registrations of users and they were able to access the system. The system efficiently authenticates the email and password, and there was successful verification of the token and CAPTCHA sent to the user's email account. The system performed excellently well meeting all the necessary objectives of the research work. The user login authentication was about 100% in accuracy and performance, the token was sent to the user email successfully and its verification is working perfectly. The captcha performance also is at its best. If the scrambled alphanumeric text is not entered properly, you will not proceed to the dashboard which is the last level to the data warehouse. Hence, the system has produced an optimal outcome which meets all-round the functionality earlier described in previous chapters.

5.2 Attempt and Frequency of Access Rejection.

The system was tried with variations of emails of different users and at different times. This done to ensure that the system is working the way it should. Below are frequency table displaying the number of times each user tries to access the system.

Table 5.1 Frequency table for wrong email and password combination

EMAIL ADDRESS	FREQUENCY
Wilzex.gmail.com	1
angelisticqueen@gmail.com	3
kantech@gmail.	1
phisher@email.com	2

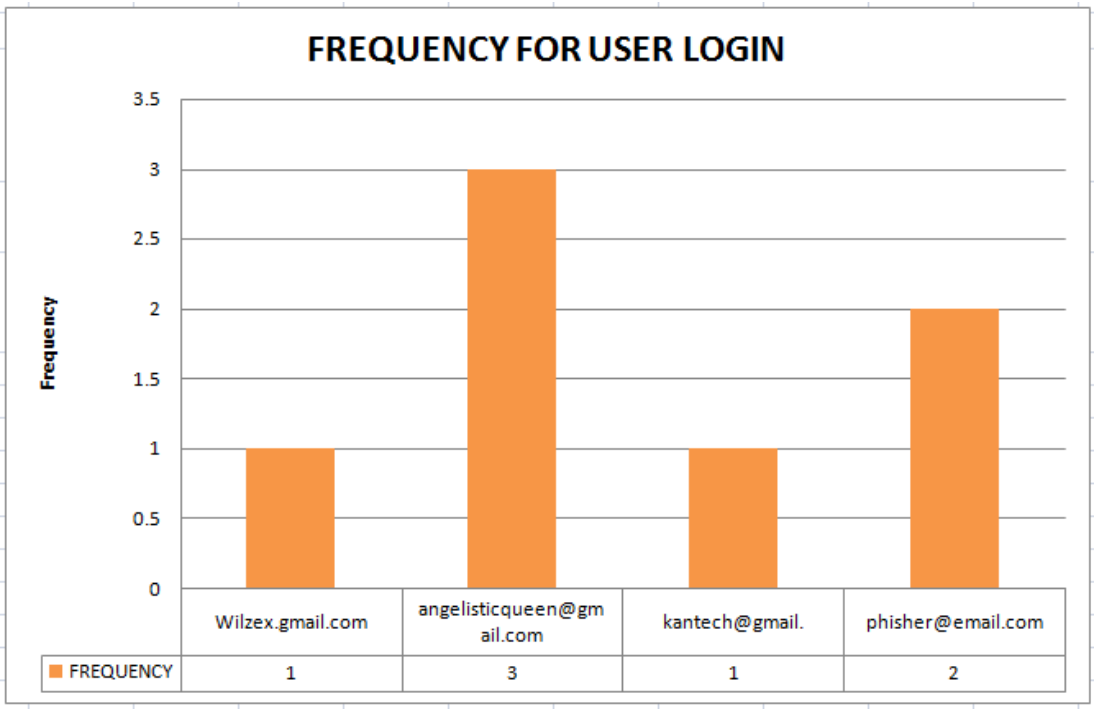

Figure 5.1 A chart showing wrong email and password combination from different users

Table 5.2 Frequency table for failed token verification

EMAIL ADDRESS	FREQUENCY
ethelscript@gmail.com	1
chromemnanger@yahoo.com	3
swizjack@gmail.com	2
maumaubless@gmail.com	2
maumaubless@gmail.com	4

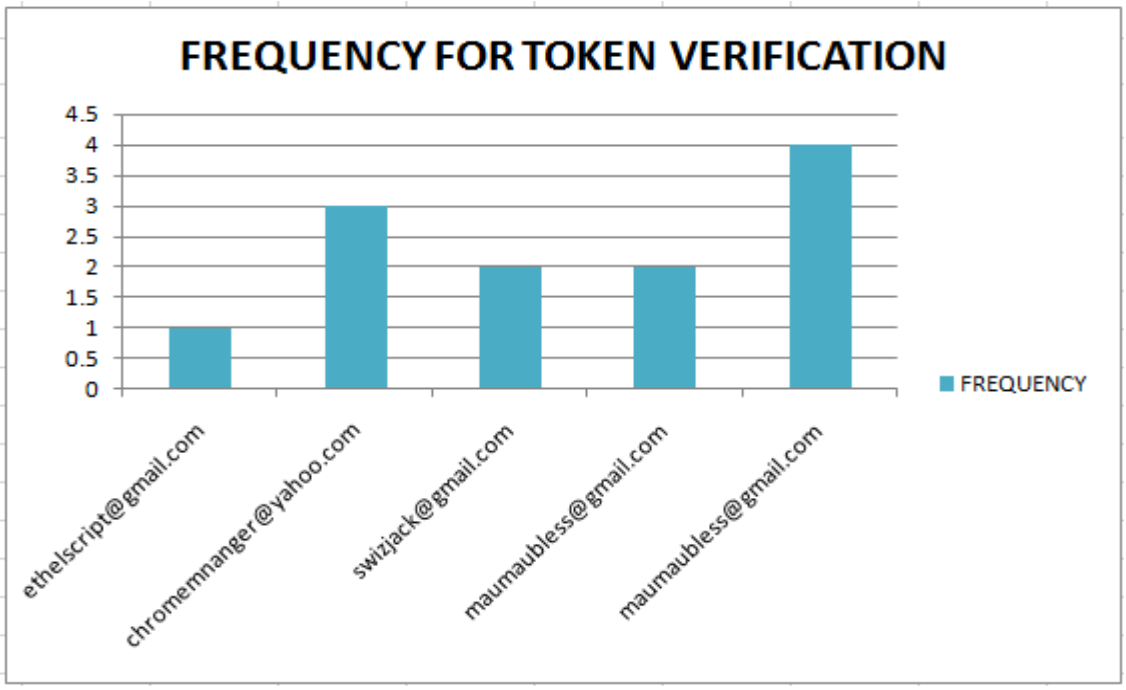

Figure 5.2 A chart showing rejected verification from different users

Table 5.3 Frequency for wrong CAPTCHA input

EMAIL ADDRESS	FREQUENCY
wilzex@gmail.com	4
Angelisticqueen95@gmail.com	6
Stargirl2011@yohoo.com	2
wilzx@yahoo.ca	1
Bridgetluv2001@yahoo.com	2
glitteringstar@ymail.com	3
Tboy519@gmail.	2

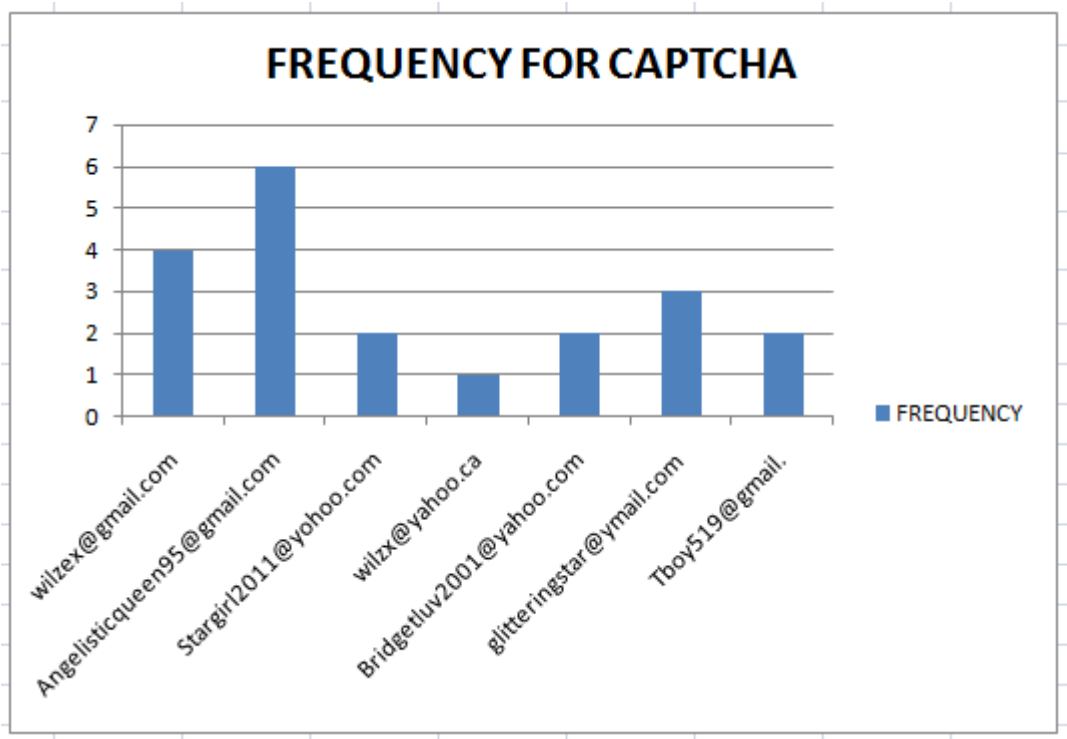

Figure 5.4: A chart showing denied access by CAPTCHA

5.4 Discussion

This system is aimed at providing a framework for Data warehouse security with authentication using a hybrid approach. The system used three levels of authentication as against two levels making the security system a unique model. For the user to be able to access the data in the warehouse, the user will have to be a registered user. The fraudulent users are not legally registered in the system, when they try to access the system they are denied access, only the duly registered user has access to the system.

The figures 4.2 to 4.4 displays forms to enter the login details, Token (One-Time Password) and CAPTCHA codes for authentication and verification of the users before access is granted. Figure 4.5 is the login access control display screen for the proposed system. When the user provides credentials that are not found in the database, the system denies access and grants access to only legitimate users. These users only have access to limited information that are related to them and won't be able to do otherwise else, denied access to the system.

The token access control is displayed in figure 4.6 the figure shows the denied access of someone that supplied wrong code and grants access to someone that supplied the right code. When a user who is not registered into the system tries to access the warehouse remotely or a robot application wants access, the token verification stops that user since the system will

require a human activity: by entering the code directly. When a wrong email (i.e. non existing email) is entered, the verification code that will be sent will not be received therefore, the user will be unable to provide the code for verification, thus, the user won't be authenticated.

The captcha access control as seen in figure 4.7 shows us an access grant to the user. The user can't proceed from this stage if the code entered into the form of the captcha screen is not correct. The scrambled alphanumeric is placed randomly on an image that makes it a little challenging for the user to decipher; Therefore causing difficulty to crack by a robot indirectly. The algorithm for this module allows the user to refresh the scrambled alphanumeric as many times as possible until he/ she can type the text correctly on the form provided.

After all security levels has been passed by the user, the system now redirect the user to his/ her dashboard to enable him/ her perform the operations permissible. Figure 4.8 shows user information gotten from the data warehouse and displayed.

The three levels of authentication during the data warehouse access checks for legal access, valid email addresses and that the user is not a robot trying to access the system. Hence, when a user provides a valid login details, the system checks for valid email address and send the user token (i.e. a one-time password) to the email address for the user to enter into the form provided. The captcha shows some scrambled text randomly placed in an image which cannot be copied or determined otherwise, except that the user type it directly into the form provided.

6. Conclusion

HMAC-Based One-Time Password and Captcha Based Authentication was used in this research. HMAC-Based One-Time Password generates Time-synchronized OTP values, based on SHA-1 based Hash Message Authentication Code (HMAC). OTP is generated and sent to the user's e-mail id. The user is then directed to next page where the user is asked to enter the OTP. The user gets the OTP using the e-mail account and enters it. If the OTP is verified the user succeeds in logging in the system. This technique has been applied to student data warehouse and been tested using student registration data. Results show that only registered students have access to the data in data warehouse. Securing data in data warehouse is of great importance and interest to small and large organization.

HMAC-Based One-Time Password and Captcha Based authentications are one of the most user friendly multi-factor authentication mechanisms that does not require an additional device. We believe our solution provides the means to secure data warehouse against brute force and other forms of attack, thus helps to prevent online account theft and fraud.

Reference

Berson, A., and Smith, J. S. (1997).Data Warehousing, Data Mining and OLAP.McGraw-hill Series on Data Warehousing arid Data Management. (45)975-988

Berson, A., and Smith, J. S. (1997).Data Warehousing, Data Mining and OLAP.McGraw-hill Series on Data Warehousing arid Data Management. (45)975-988

Blaxter L., Hughes C., and Tight M. (2006). How to research. Open University Press, Maidenhead, Berks.

Caserta, J., and Kimball, R. (1998). The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publication , 5(1), 37-48..

Dedić, N. and Stanier C., (2016)., "An Evaluation of the Challenges of Multilingualism in Data Warehouse Development" in 18th International Conference on Enterprise Information Systems 6(189- 196.

David M. K and David J. A. (2015): Database Processing, Information System Relationships in 18th International Conference on Enterprise Information Systems 6(9) 189- 196.

Inmon, W. H. (2006). Data warehouse 2.0-Architecture for the next generation of data warehousing. DMReview.DM Direct Nesiiletter. . *ISBN 0-13-063085-3,261-270.*

Inmon, W. H. (1991). Building the Data Warehouse.Wiley and Sous. . *ISBN 2-5-167085-3,61-70.*

Kimball, R. (1997). Hackers, Crackers, and Spook, Ensuring that yuurdata warehouse is secure. In Journal DBMs 10(1)14-30

Mallah E.G. (2000)."Decision support and Data warehouse System" Tata McGraw-hill. 1(1)14-30

Okerenke H. (2015). The information flow in Data Warehouse ETL Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publication , 5(1), 37-48..

Santos, R., Bernardino, J., and Vieira, M. (2011). A survey on data security in data warehousing: challenges and opportunities. EUROCON - International Conference on Computer as a Tool (EUROCON), pp.1-4. IEEE. 5(1)14-30

Carlin, K. E., Russo, D. R., & Balfour, B. (1990, December). A proposal for a recursive object-oriented life-cycle. In *Proceedings of the conference on TRI-ADA'90* (pp. 156-167). ACM.

Santos, R. J., Rasteiro, D., Bernardino, J., & Vieira, M. (2013, April). A specific encryption solution for data warehouses. In *International Conference on Database Systems for Advanced Applications* (pp. 84-98). Springer, Berlin, Heidelberg.

Yeow, W. L., Mahmud, R., & Raj, R. G. (2014). An application of case-based reasoning with machine learning for forensic autopsy. *Expert Systems with Applications*, *41*(7), 3497-3505.

Thangaraju, G., & Rani, X. A. K. (2016). A Survey on Current Security Perspectives in Data warehouses. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, *19*(2).

Reddy, M. S., Reddy, M. R., Viswanath, R., Chalam, G. V., Laxmi, R., & Rizwan, M. A. (2011). A Schematic Technique Using Data type Preserving Encryption to Boost Data Warehouse Security. *International Journal of Computer Science Issues (IJCSI)*, *8*(1), 460.