# A Comprehensive Security Model for Patient Data Warehouse

**Paramasivam. Suraj thyagarajan** [1], **Neeraj Harikrishnan Girija Paramasivam** [2]
[1]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[2]Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, MI, USA

**Abstract**— *This paper deals with implementation of a three pronged security model for data warehouse, which is comprehensive and fool proof. The three pronged model implements security at the ETL level, by introducing aggregation and data masking and establishes database level security by implementing the virtual private database. This model prevents any direct access to the actual data warehouse, and all accesses are provided only through the VPD. This ensures that the database is secured from unauthorized access. The standard methods of data masking and aggregation are used to prevent any attacks on the data warehouse. The combined power of these methods, give a comprehensive security model, that can be implemented on enterprise wide or public data warehouse.*

**Keywords:** Datawarehouse Security, ETL Security, Virtual Private Databases, Three pronged security to Datawarehouses, patient data security

## 1. Introduction

In modern day data warehouses, the implementation of security has not been given a great deal of importance and this sometimes leads to data theft at various levels. This project aims at preventing the data theft from the data warehouses. The project proposes a three pronged approach to implement security in datawarehouses. We also discuss the other related work in this area, distinguishing them from the current work. This paper looks into a real time experiment conducted on a large dataset of patient genomic data and compares the performance parameters with a regular system.

A datawarehouse is a large collection of data in a large specialized database. These datawarehouses are used for a variety of purposes including data mining, business intelligence and many more. The databases generally used for these purposes are large scale machines with high reliability and scalability. The best examples of such specialized systems are Oracle Warehouses, Teradata, and netezza systems. However, these systems follow security through obscurity principle.That is they consider the system is secure because limited users have access to it. The security models employed for most of the corporate data warehouses are not comprehensive in providing security.

In terms of security implementation, the process of data loading, defined by ETL, has a lower security priority for most organizations. The process of ETL is defined as Extraction, Transformation and Loading, which is the process of loading data into the Datawarehouses. Extraction refers to the process of extracting data from multiple sources. Transformation refers to the process of cleaning data or applying business rules to the data. Loading is the actual process of loading data into the warehouse. There are other kinds of processes which are used to load data into the warehouse, which include ELT(Extract Load and Transform) etc. However, none of these loading techniques seem to implement data security at the ETL level albeit there are a few works suggesting security at this level. However, these papers do not propose a model that encompasses both database and ETL security.

Another level of security that is implemented by some of the datawarehouses is the provision of controlled access to the DW users. The access restriction is usually based on a simple authentication mechanism, which authenticates users based on their credentials. However, most data warehouse set ups do not have rigorous implementations of authentication mechanisms for these. This provides with a small amount of security to the databases, which unfortunately can be broken with not too much effort.

This paper aims at combining some of these well known security practices to put forward a comprehensive model that aims to secure a datawarehouse in the health sector. The issue of security in the health related data is one of the most important, yet one of the most overlooked of all. This paper aims in providing comprehensive security to a patient genomic datawarehouse. However, our model can be implemented across any domain and can be extended to any generic data warehouse.

## 2. The Three Pronged Model

We here propose a three pronged approach to security, for providing comprehensive security for the patient datawarehouses. This is a generic model that can be applied to any data warehouse per se. However, we have handled patient datawarehouse as a case. We also have presented an analysis of the performance of the loading process of data. We can safely presume with numerous examples that the performance of a warehouse is more dictated by the ETL process rather than by any downstream processes.

In this model, we look at three different areas where the data resides or traverses. We divide these into high risk areas, low risk areas and presumably safe areas. In the figure [1] , we provide the complete overview of the

model that has been devised. In the figure [1], the first area, containing the sources and the staging area, are considered as the safe areas. These areas have least security threat due to the corporate nature of their maintenance. These areas where data reside can hence be presumed to be safe. The next important area where the data resides and traverses is the public datawarehouse area, which is prone to attacks of various kinds. However we can further subdivide this area into smaller fragments which have different risks. The target data warehouse is where the data is originally housed. This area is prone to different kinds of attacks. The data in this area is to be protected since this is the final set of data, that contains sensitive information. To mitigate the risks in this area, we suggest a secure approach, in the way data is being populated into this database. The principles of aggregation and generalization is well known in the industry for two purposes. The concepts of aggregation is used both for improvements in the overall security of the warehouse and also for improving performance of the joins during mining of data. However, we propose to use aggregation in conjunction with data masking, to provide additional security to the warehouse. Aggregating data will let us hide some parts of data from anyone using the data warehouse. Also the usage of data masking obfuscates the most important columns from any one accessing the warehouse. However, the aggregate tables and data masking transformations together can provide improved security to the entire set up. The aggregate tables are set up in such a way that only the VPD engine can access it. The same applies to the Data Masking tables. This provides enhanced security,since these tables contain vital information which can be misused when joined with the data from the actual data warehouse.

We also introduce the concept of virtual private database in this context, making it the only engine which will have access to both the aggregate and data masking tables along with the actual warehouse tables. This makes it impossible for any intruder to steal any meaningful data from the warehouse. The virtual private database is a concept prevailing for a long time and is one, which gives different views of data to different users based on their roles and privileges. We, in our experiment have used an oracle virtual database on a patient genomic data system.

In order to ensure user authentication, most of the existing systems use built in security mechanisms of the databases or other Business intelligence or data mining tools. However these tools are sometimes not very efficient in providing ample amounts of security and can be easily broken. To mitigate this, we propose to use a separate authentication server, which handles a DES keystore, to authenticate users. This authentication will be a two way authentication and will provide a more fool proof method for user authentication.

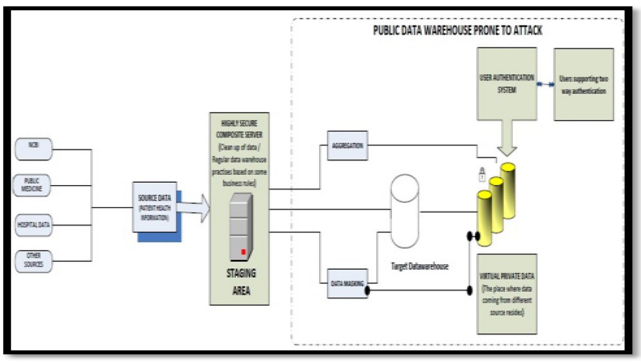


Fig. 1: The block diagram for the three pronged model. Refer Fig 1 in Appendix for a bigger picture

## 3. A Comparison with some Related Work

Though there are several related works in the field of database security, the implementations of security to ETL is studied in a very few works. One such work was presented at the Third UK Symposium on Computer Modelling and simulation by M Mrunalini et all [1]. The authors in this piece of work look at implementing ETL security and also provide appropriate UML based use cases for a model to implement ETL security. However, the authors do not concentrate on any of the database security options. Also the paper deals with the methods of user authentication and implementing security, which are generic in nature and are applied to the principles of datawarehousing. In our work, we have developed a specific model for ETL , which are based on the well known concepts of ETL. Also we here provide a more comprehensive model, which includes database security and ETL security within the same model, along with user authentication systems. This makes our system unique with a more comprehensive nature. There are other related material available on the internet and other sources, which talk about security in ETL or security in databases. Also the use of virtual private databases is being extensively referenced by various literature. However none of these material seem to suggest a comprehensive model for security of data through the entire process of data warehousing. A virtual private database for example, is usually not accompanied by any security measures at the ETL level. The security at ETL level is ignored by most authors, since they assume ETL to be inherently secure and restricted to a corporate sector. However, due to increasing threats to corporate data, which includes sensitive data, by insiders, the need to implement ETL security has gained extreme importance. Another related work by Kimmo Palletvuori from Helsinki University of technology titled "Security of Data Warehousing server" [2] details the methods of implementing security to the physical servers for the datawarehouse. This deals more with the security of the server at a physical level rather

than at a logical level. The author in this paper also talks about concealing data at a logical level and preventing unauthorized access at the ETL level. However, this paper does not detail any methods of providing security at the ETL level. This makes the process less secure by not providing security for the entire process. Our work in this project provides a method to mitigate this gap in providing security to the data in the data warehouse. The project gives a method of providing ETL security and security to the datawarehouse, using industry standard methods. This project details a comprehensive method of providing data security to the data warehouse. Another major work related to the area of datawarehouse security is the white paper presented by Oracle in April 2005 [3]. This white paper provides complete details of the virtual private database that oracle is capable of providing. However, this white paper also does not provide any details on the possible ETL level security. This provides a gap in the security arrangement provided inherently by oracle. This differentiates our work from the existing white paper published by oracle. Our project provides a more comprehensive model which fills some of these existing gaps in providing security to the data warehouse.

# 4. Experiments and Results

Table 1: Table indicating the run times using various ETL tools.

| Tool Used | Number of Records | Run Time without Security features implemented | Run Time with Security features. |
|---|---|---|---|
| Informatica | 11104 | 3.00 minutes | 4.02 minutes |
| Informatica | 4509237 | 15.08 minutes | 20.43 minutes |
| TalenD | 11104 | 3.24 minutes | 4.29 minutes |
| TalenD | 4509237 | 16.23 minutes | 23.07 mins |

This project was carried out on a patient data warehousing system, with industry standard databases and ETL tools. The project was carried out on a large dataset obtained from the NCBI. Due to privacy concerns, we created hypothetical data using the genomic accessions available in the NCBI database. We created a source dataset, with hypothetical names and SSNs, which did not reflect actual data, however reflected the volume of data found typically on a patient datawarehouse. Additional columns were introduced to create a more logical dataset, and this dataset was loaded into the data warehouse. Two sets of loading were performed on the data warehouse. The first set of loading was done without any security implemented on ETL or the database. The second set of loading was done after implementing the security model on the ETL and databases. This provided us a clear picture of the performance degradation on the entire process. In this case, we used a dataset of 11104 records in the source data. The first run was conducted and this took an overall run time of 3 minutes to load the entire set of data into the warehouse. The second run, with the security features implemented took a total time of 4 minutes and 02 seconds. This was achieved with a total throughput of 1110 records using an industry standard tool for ETL, Informatica. Also all these runs were conducted on Oracle 10g and a machine with Intel Core 2 Duo processor running with a RAM of 2 GB. We also performed the same experiment with an open source ETL tool, TalenD, to ascertain the results were not biased by the tool and we could get accurate results. Below given is a table of values obtained during the experiments using both the ETL tools.

The entire experiment showed that the degradation in performance due to the additional implementation of security was very minimal and hence this model of implementation of security can be implemented without too much overhead on the entire system. The screenshots of the execution of our project is available in the appendix.

# 5. Conclusion

The project has proposed a means of implementing security using a new model, which has been proposed to fill any gaps in security of data warehouse. The notion of security in data warehouses have so far been restricted to security in databases. A very few related works have been related to security in ETL. However, none of these models are comprehensive, providing both ETL and database level security. This project provides a total security for the entire data warehouse along with the process of loading data using ETL. The scope of future work in this project can be extended to creating models implementing cryptographic protocols at the ETL level, which can provide more efficient security. The framework we provided here is open and can be extended to different levels or different domains. However, this existing domain of health care and security of patient data is extremely important and it is important to provide more secure ways of data access in these data warehouses. Also the work could be extended to provide a model of security for the data mining paradigm.

# References


[1] M Mrunalini, T V Suresh Kumar, K Rajani Kant, "Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes," *Third UKSim European Symposium on Computer Modeling and Simulation*, 2009.

[2] Kimmo Palletvuori, " Security of Data Warehousing Server,"

[3] *Security and the Data Warehouse, An Oracle White Paper*,2005.

[4] Wiley Publications 2006 *Ralph Kimball, The Datawarehouse ETL Toolkit*.

[5] A. Simitsis, P. Vassiliadis, and T. K. Sellis, "Optimizing ETL processes in data warehouses,"*In Proc. ICDE, pages 564-575*, 2005.

[6] D.W. Embley, D.M.Campbell, Y.S.Jiang, S.W. Liddle,D.Wlonsdale, Y.-K.Ng, and R.D. Smith,"Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," *Data and Knowledge Engineering 31,no. 3* pp. 227-251 Nov 1999

[7] S. Jajodia and D. Wijesekera, "Securing OLAP data cubes against privacy breaches," *In Proc. IEEE Symp. on Security and Privacy* pages 161-178 2004.

[8] A. Simitsis, " Mapping conceptual to logical models for ETL processes," *In Proc. DOLAP*, pages 67-76, 2005

[9] T. Priebe and G. Pernul,". A pragmatic approach to conceptual modeling of OLAP security," *In Proc. ER, pages 311-324* 2000

# 6. Appendix

**The Architecture Block Diagram**

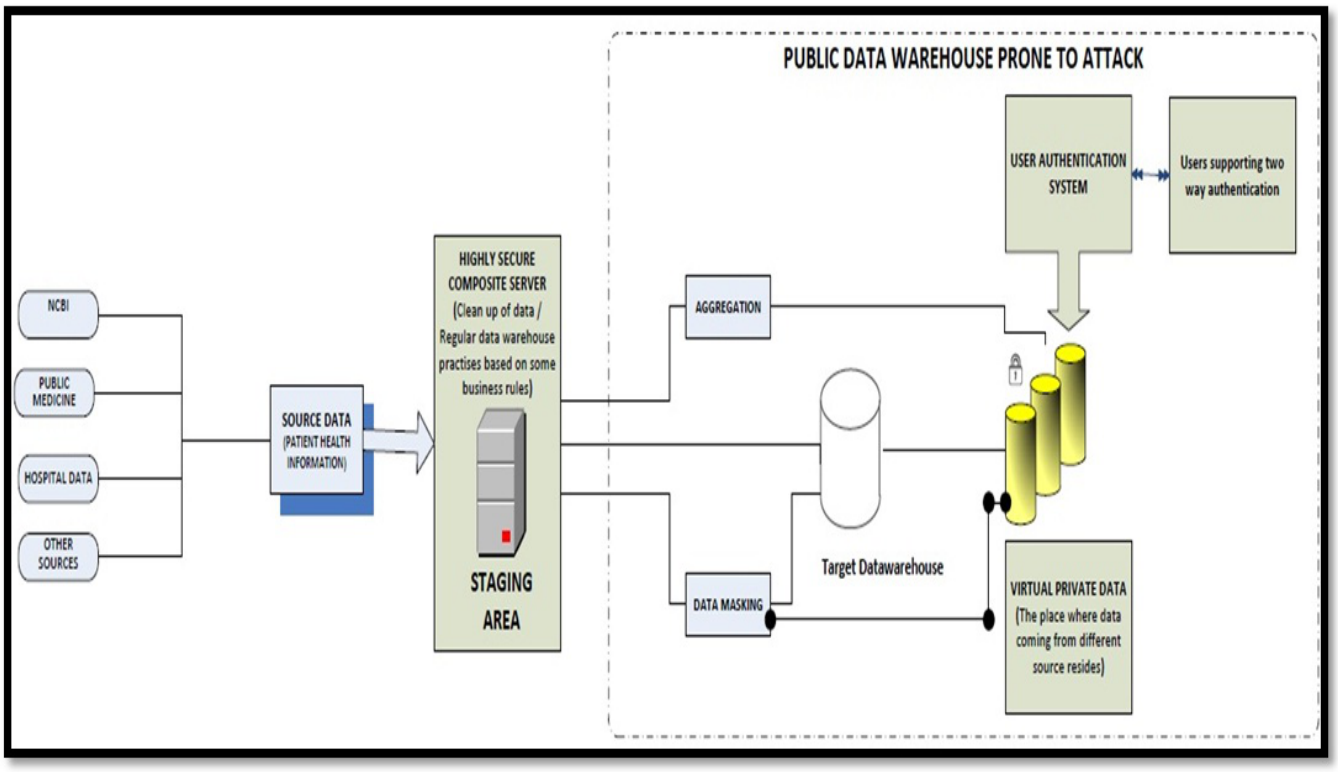


Fig. 2: Enlarged Version of architecture.

**Screenshot of Experiment Results:**

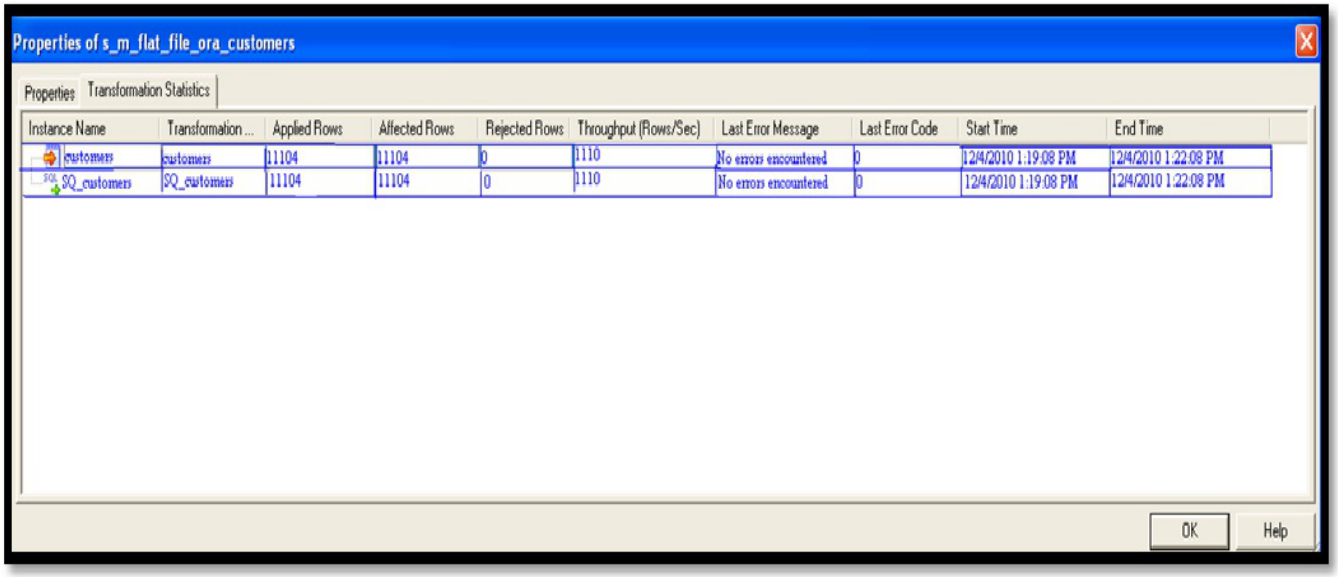


Fig. 3: The result of experiment on Informatica showing the time required to load 11104 records without security module implemented

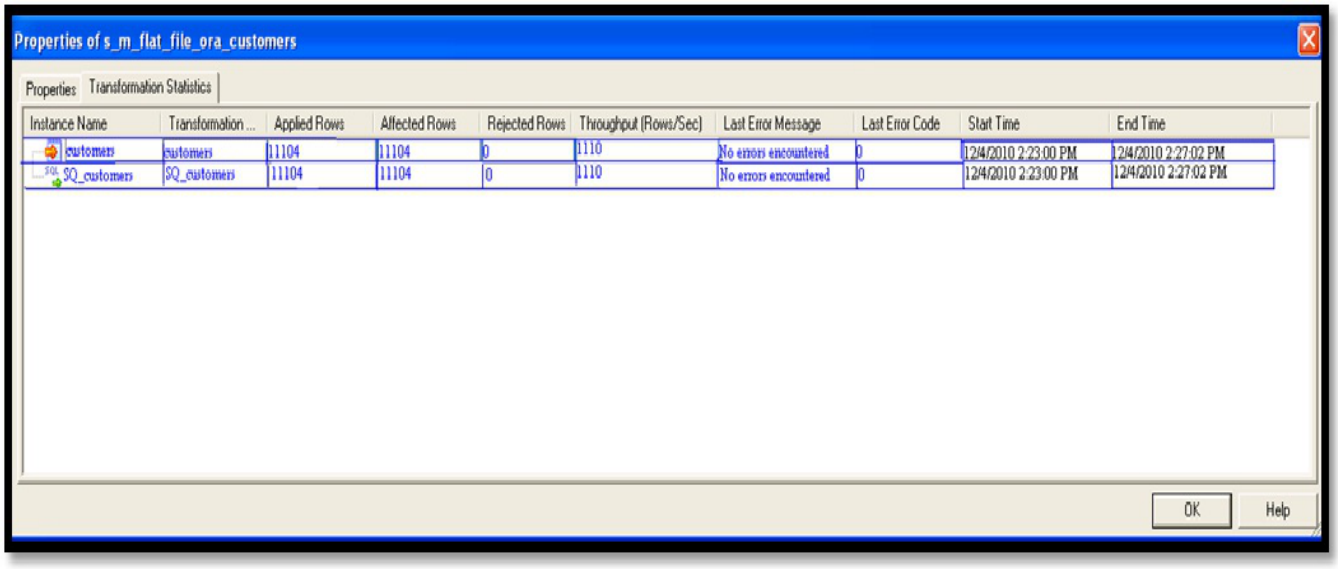


Fig. 4: The result of experiment on Informatica showing time required to load 11104 records after implementing security model