

Security Measures for Web ETL Processes

Salma Dammak, Faiza Ghozzi Jedidi and Faiez Gargouri

Abstract Security aspects currently play a vital role in software systems. As security managers have to operate within limited budgets they also have to patch up the increasing number of software security vulnerabilities. They need to perform a risk evaluation in order to determine the priority of patching-up vulnerabilities. The use of quantitative security assessment methods enables efficient prioritization of security efforts and investments to mitigate the discovered vulnerabilities and thus an opportunity to lower expected losses. Elsewhere, Extraction Transformation Load (ETL) processes, known as a core, development of WeBhouse. Securing these processes is highly important and helps in mitigating security defects in decisional system. For this purposes, this paper adopts the Common Vulnerability Scoring System (CVSS) and proposes a Meta model for security measure in Web ETL processes enabling security manager to asset anticipated vulnerabilities.

1 Introduction

WeBhouse represents a new form of data warehouse that collects data from the web, which presents a large source of information, in order to integrate the data from several sources. Due to sensitive data contained in the DW, it is important to assure the security of the decisional system from the early stages of the life cycle of its development. In the Web context, the mega-volumes of data, the increasing number of users and the heterogeneity of requirements increase new challenges, as the nature of data, the lack of control over the sources, and the frequency of changes on data.

S. Dammak (✉) · F.G. Jedidi · F. Gargouri
MIRACL-ISIMS Pole technologique de SFAX BP 242-3021,
Sakiet Ezzit Sfax, Sfax, Tunisia
e-mail: damak.salma@gmail.com

F.G. Jedidi
e-mail: jedidi.faiza@gmail.com

F. Gargouri
e-mail: faiez.gargouri@gmail.com

Besides, the main challenge in this field is how to integrate heterogeneous and safety data in the WeBhouse when some or most data sources reside on the Web.

Indeed, the use of Internet in a storage process (acquisition, storage and access) increases the risk of attacks that can affect decisional system and make it more vulnerable. An improper disclosure of such data represents a risk to the system. These drawbacks lead us to consider that security is a key factor for a WeBhouse. In addition, the security needs are not considered at the earliest designing phases of data warehouses; also most design methods do not integrate them. An engineer who does not have a complete description about security requirements and the way of securing data will treat these aspects at the closing stages of the WeBhouse development.

Note that the Extraction-transformation-loading (ETL) processes play an important role in WeBhouse architecture because they are responsible for integrating data from heterogeneous data sources. But, works dealing with ETL processes security are few and far from proposing a complete approach to secure the ETL web process. They are focused mostly on the extraction process although the transformation process covers major complex operations. Also, the loading process requires a protected network to charge data in the WeBhouse. It is therefore widely recognized that the appropriate design and security of the ETL processes are key factors in the success of WeBhouse projects.

At the design and implementation stages of WebHouse life cycle, vulnerabilities are inadvertently introduced and result from defects or weaknesses. They can be exploited by attackers to harm the system. Therefore, security needs to be considered and measured from the early stage of the development life cycle. To evaluate the security and safety level in WEB ETL processes, we define a secure Meta model. Our meta enriches the Meta-model of UML activity diagram defined by OMG [1] by adding security classes based on the metric groups of the CVSS (Common Vulnerability Security Score) [2]. Our approach enables us to verify the safety of the data source before the extraction process, to detect vulnerability and define for each one the suitable preventions and solutions and to secure the inter-phase transition. The proposed solutions integrating security needs are melted in the CVSS factor metric groups values, thus enabling to asset the severity impact of vulnerabilities. This measuring allows us to prioritize vulnerability and reduces the impact of vulnerabilities on the ETL process because the most severe will be the first remedy. Our proposition is applied to a case study for the Web ETL processes of an online banking system, due of the risk that may come across this system and the sensitivity of the data that they process.

This paper is organized as follows: Sect. 2 surveys ETL process security and security measurement related works. Section 3 presents our secure Web-ETL Meta-model integrating vulnerabilities specification. Section 4 depicts our secure web-ETL case study instantiating our meta model in banking WeBhouse ETL processes. Section 5 concludes the paper with some pointers to future directions.

2 Related Work

2.1 *ETL Process and Security*

2.1.1 Multidimensional Integration of Web Data

The proposed Web Warehouse architecture in [3] is composed of two parts: a repository of functional characteristics of a data warehouse and a repository of Web click-streams describing the data. Before data extraction, the approach requires the verification of the accuracy of data sources from the Web. Web logs are unified in [4] the proposed approach is to represent the raw data log file in a model that conforms to a meta-model of generic Web log. Once this model is designed, a multidimensional conceptual model is obtained automatically through the transformation language QVT (Query/View/Transformation). Based on the SOA architecture, [5] discusses the real-time data analysis and they design the real time data warehouse architecture. In the proposed architecture, a multilevel real-time data cache is introduced to facilitate the realtime data storage. Reference [5] combines the Web services and data warehouse by using the Web service to capture real-time data and make the real-time data capture and load easily.

However, to summarize the current literature, we note that all the work deal with Web inputs. We find that the treatment of ETL processes which present an important task for designing the Webhouse is infrequent or absent. In addition, non-functional security requirements are not addressed in any design approaches with the exception of work [6] which has emphasized the need to verify the accuracy of data sources. Therefore, in the next section, we present the few works studying the safety of ETL processes.

2.1.2 ETL and Security

Although there are conceptual proposals for the modeling of ETL processes, there are few proposals that support safety issues. The complexity of the ETL system arises as the sources use different data formats, which has to be cleaned, transformed, aggregated and loaded into the data warehouse as homogeneous data. Consequently and with the emergence of the web, different security problems are triggered and a security approach for the ETL process becomes a urgent need. Reference [7] proposes a simulation model for the extraction of secure data during the ETL process. To accomplish the conceptual modeling of Secure Data Extraction in ETL processes, they extend the UML uses case and sequence diagrams. An algorithm is proposed based on the users authentication, the encryption of the extracted data, the verification of data corruption. This work is extended in [8] where authors analyze the ETL process under two security metrics: vulnerability index and security index. A framework for securing any phase in the ETL process has been suggested together with a methodology for assessing the security of the system in the early stages. Reference

[8] develops a simulation tool SeQuanT which quantifies the security of the system, in a general context. An object-oriented approach to model the ETL process embedding privacy preservation is proposed in [9] the researchers present a class diagram for ETL embedding Privacy Preservation. The main components of this diagram are Data Source, retrieval component, source identifier, join, PPA and DSA. The PPA is the most important and major component of this architecture, its task includes cleaning and transforming data followed by privacy preserving attributes.

There are few research papers on ETL security. These researches do not take into account non-functional security needs of ETL processes in the Web business level and they do not address the vulnerabilities between processes of Web ETL.

2.2 Security Measurement

To secure a system, we must identify and assess vulnerabilities. So we need to prioritize these vulnerabilities and remediate those that pose the greatest risk. Several research papers addressing topics of measuring information security have been written in the last few years. National Institute of Standards and Technology have published Special Publications dealing with topics of Risk Management and Information Security Metrics [10–12].

Several studies are based on the CVSS which is an adopted standard that enables security analysts to assign numerical scores to vulnerabilities according to severity to measure security [13]. In [14], authors propose the novel approaches of handling dependency relationships at the base metric level, and aggregate CVSS metrics from different aspects. A CVSS risk estimation model is presented in [15]. The proposed model estimates a security risk level from vulnerability information as a combination of frequency and impact estimates derived from the CVSS.

3 The Meta Model Proposed for Secure WEB ETL

To model our secure Web ETL processes, we extend the Web ETL meta model presented in [16], based on the activity diagram Meta model UML OMG [1] and we enrich it with our security Meta classes. Because we cannot improve what we cannot measure, [17, 18], measuring security in Web ETL process is an essential. By applying security metrics, we can judge the relative effectiveness of our proposal [13]. The Common Vulnerability Scoring System (CVSS) is an adopted standard that enables security analysts to assign numerical scores to vulnerabilities according to severity [13]. It calculates the vulnerability score using attributes grouped into three metric groups: base, temporal and environmental. A CVSS score is a decimal number in the range (0.0, 10.0), where the value 0.0 has no rating (there is no possibility to exploit vulnerability) and the value 10.0 has full score (vulnerability easy to exploit). The base metric group (mandatory) quantifies the intrinsic characteristics

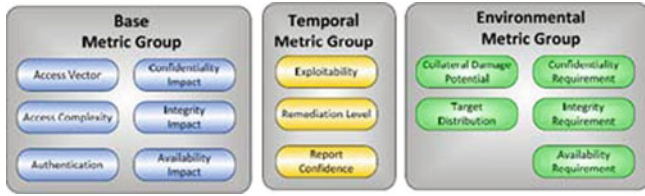


Fig. 1 CVSS metrics groups

of vulnerability in terms of two sub-scores: (i) exploitability subscore; composed of access vector (AV), access complexity (AC) and authentication instances (Au), and (ii) impact subscore to confidentiality (C), integrity (I) and availability (A) [14] (Fig. 1).

Scoring equations and algorithms for the base, temporal and environmental metric groups are described below. The base equation is the foundation of CVSS scoring. The base equation is:

$$BaseScore = ((0.6 \times Impact) + (0.4 \times E)1.5) * f(Impact) \quad (1)$$

With:

$$Impact = 10.41 \times (1 - (1 - C) \times (1 - I) \times (1 - A))$$

$$Exploitability = 20 \times AV \times AC \times Au$$

$$f(impact) \begin{cases} = 0 & \text{if } impact = 0 \\ = 1.176 & \text{otherwise} \end{cases}$$

The temporal (optional) equation will combine the temporal metrics with the base score to produce a temporal score ranging from 0 to 10. Further, the temporal score will produce a temporal score not higher than the base score and it is composed of Exploitability (E), Remediation Level (RL) and Report Confidence (RC). The temporal equation is:

$$TemporalScore = (BaseScore \times E \times RL \times RC) \quad (2)$$

The environmental (optional) equation will combine the environmental metrics with the temporal score to produce an environmental score ranging from 0 to 10. The environmental equation is with CDP (CollateralDamagePotential), TD (TargetDistribution), (CR/IR/AR) security requirement respective to Confidentiality/Integrity/Availability.

$$EnvironmentalScore = ((AT + (10 - AT) \times CDP) * TD) \quad (3)$$

With:

AdjustedTempora(AT) = TemporalScore recomputed with the Base Scores
Impact subequation replaced with the AdjustedImpact equation

$$AdjustedImpact = min(10, 10.41 \times (1 - (1 - C \times CR) \times (1 - I \times IR) \times (1 - A \times AR)))$$

In our Meta model, we represent the Environmental Metric Group in a Meta Class while the Temporal Metric Group factors are represented in the vulnerability catalog. Some of the Base Metric Group factors are presented as Meta class like Access, Security Aspect, others are presented as Meta class association like the Authentication, the Access Complexity.

In order to consolidate the anticipated vulnerabilities in the WEB ETL process, we define a Metaclass VulnerabilityCatalog. This Meta class contains the most critical vulnerability that can be detected during the three WEB ETL processes (Extraction, Transformation, Loading) and we take into account the vulnerabilities encountered between processes. The meta class Vulnerability inherits all the Vulnerability Catalog attributes and has the state attribute having values: detected, corrected, expected. This categorization minimizes the probability that vulnerability happens by proposing preventive actions and enables to define a secure environment. Also, it accelerates the remediation actions since the security manager knows from the beginning the most of the vulnerabilities that can be detected and he has already prepared the appropriate remediation solutions.

As mentioned previously, our Meta model is composed of two packages: the Web ETL Meta model and the security Meta model. In the following part, we describe the Web ETL Meta model represented in Fig. 2:

- Steps inherit from “ActivityGroup” class. It models the extraction, transformation and loading processes. Each “Step” is composed of “Activities”.
- Sons and Parents inherit respectively from the “inputPin” and the “outputPin” classes.
- Xml Buffer, Log file, Website and Relational DB inherit from the DataStoreNode class are used as storage means.
- XML Dimensional is an XMLBuffer class. It represents a dimensional structure (fact, dimension, attributes ...).

The security Meta Model is presented with blue color in Fig. 3 and described as follow:

- Security Need represent user requirements that must be satisfied in our security approach.

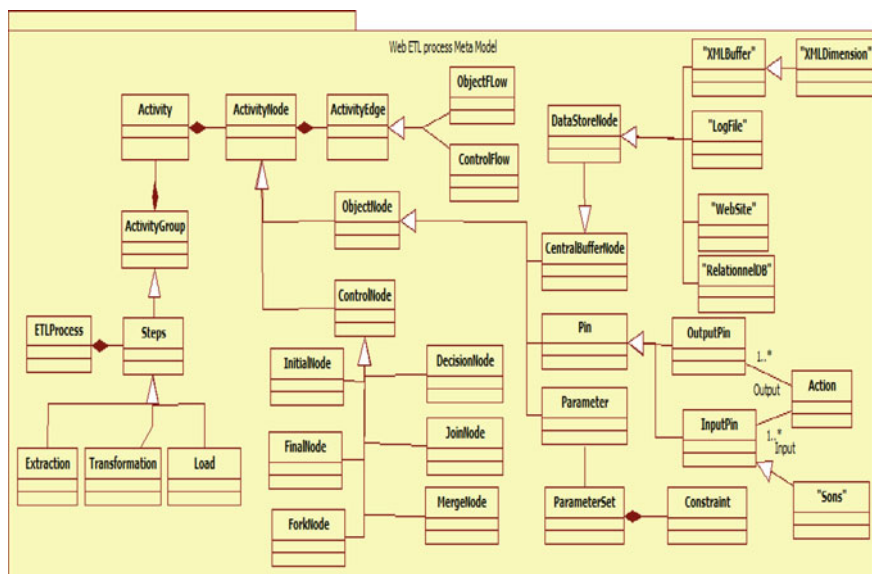


Fig. 2 Web ETL meta model

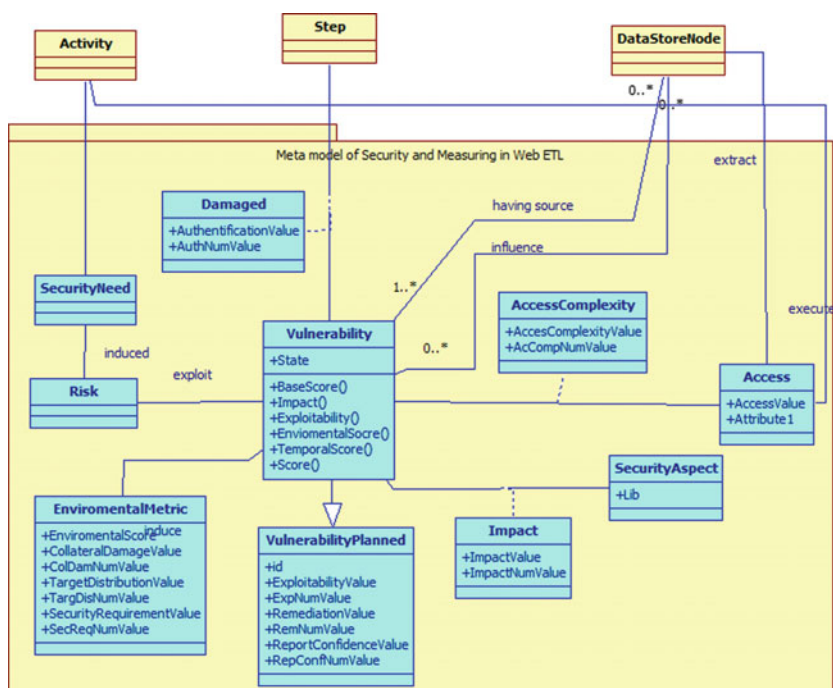


Fig. 3 Web ETL meta model

- Risk is security problems preventing the achievement of security goals.
- Vulnerability Catalog is the anticipated vulnerability having as attribute the three factors of the temporal Metric.
- Group: confirmation of the technical details of vulnerability, the remediation status of vulnerability, and the availability of exploiting code or techniques. Since temporal metrics are optional, they include a metric value that has no effect on the score. This value is used when the user feels that the particular metric does not apply and wishes to skip over it.
- inherits from the Vulnerability Catalog Meta class.
- Access present the type of access to the data (local, adjacent network).
- Security Aspect are Confidentiality, Integrity, Availability.
- Environmental Metric Group captures the characteristics of vulnerability that are associated with a users IT environment.
- Impact is an association class between the Security Aspect and the Vulnerability and measures the impact value.
- Affect is an association class between the Access and the Vulnerability and measures the Access Complexity.
- Damage is an association class between the Steps and the Vulnerability and presents the authentication metric.

4 Scoring Security Measure in Banking WeBhouse ETL Processes

Our proposition is applied to secure the Web ETL processes of a banking system. We focus on each process activity and define for it vulnerabilities that can affect quality and safety of data. To achieve this purpose, our Meta model, covering ETL process and CVSS vulnerabilities, is instancied. The first step in our work is the definition of non-functional requirements depicted by SecurityNeed. Next, we search and set, for the defined requirement the major risks that threaten the meeting of these needs. This work leads us to set vulnerabilities and categorize them. So, for each vulnerability we identify the affected activity during the process (Extract, transform, load), the access type, the security aspect that it affects, the environmental and temporal factor. All these parameters enable us to calculate the CVSS score and to set the safe remediation solution.

4.1 *Extraction Process*

At the Extraction phase, some specific activities are defined to integrating several web sources, especially log files and websites [16]. This step consists of extracting data from log files (stored in web servers), websites and sources (relational DB, XML ...).

Table 1 Security risks in the WEB extraction process

Risk	Vulnerability
Improper input	Violation of customer private information
	Erroneous and incorrect URL data
	Entrusted Website
	Entrusted data (virus, malware)
	Lack of integrity controls
	Missing encryption of sensitive input
Administrator identity problem	Unprotected transport credentials
	Weak cryptography of password
	Insufficient verification of data authenticity
	Improper autorisation
	Lack of access control
	Cybercrime: Phishing

The main objective of this step is to unify these sources, into a single format to be used later. Our task consists in controlling the safety of source and to secure the extracted data. In Table 1 presents the majority of security problems intended in the Extraction phase. Two risks are defined and for each one the corresponding vulnerabilities are specified.

In the next paragraph, we will give an example of how to present security in the activity diagram ‘Log file structuring’. Structuring log file is an essential activity that allows structuring and organizing a log file in an XML form. We study two vulnerabilities which can be detected in this step and calculate the CVSS score.

Unprotected Transport of Credentials

Figure 4 presents ‘LogFileStructure’ activity describing the log files structuring process enriched by the security problem that can decrease the process quality. To measure the vulnerability Unprotected transport of credentials, we set values for each factor of the three metric groups of CVSS: base metric, environmental metric and temporal metric. For the base metric, this vulnerability affects greatly (AC: 0.35) the access by an AdjacentNetwork (AT: 0.646).). It does not damage the authenticity (Au: 0). It affects completely Confidentiality (C: 0.660), and has no impact on Integrity (I: 0) and availability (A: 0).

As an environmental factor, the detected vulnerability increases the collateral damage (CDP: 0.5) and the target distribution (TD: 1). It affects security requirements by increasing confidentiality and reducing integrity and availability ([CR/IR/AR], [1.51/0.5/0.5]). For the temporal factor, this vulnerability has a Proof of concept Exploitation value (E: 0.99) i.e. the code or technique is not functional in all situations and may require substantial modification by a skilled attacker. It has a Workaround Remediation Level (RL: 0.95) i.e. there is an unofficial solution available and it has an Uncorroborated Report Confidence (RE: 0.95). Based on the defined factors value in Fig. 4, we calculate the Base, Temporal and Environmental score.

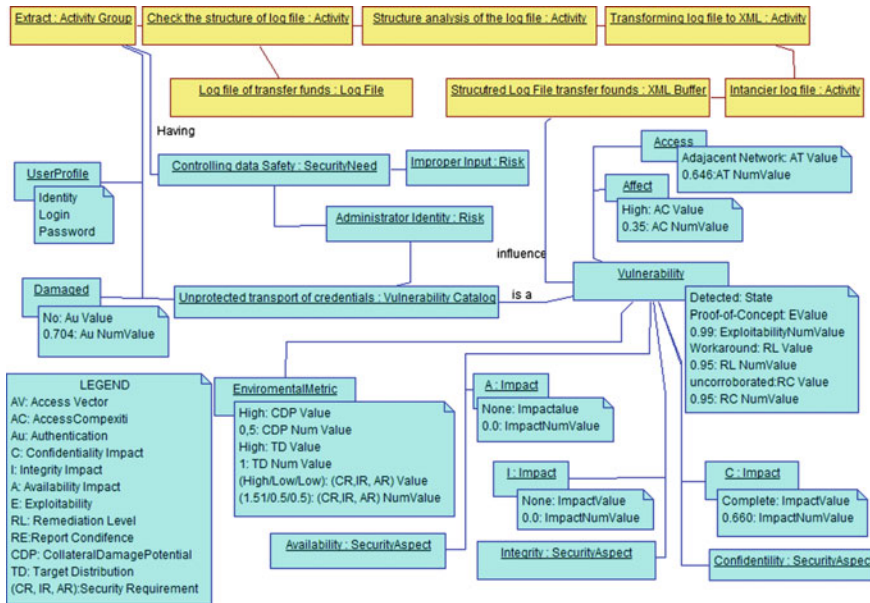


Fig. 4 ‘Unprotected transport of credentials’ security factors in the “Log File Structure” activity diagram

BaseScore = 4,58
 Temporal Score = 3,72
 Environmental Score = 2,93

Violation of Customer Private Information

In Fig. 5, we present the same activity but the security need protection of the authentication parameters is treated. The CVSS score are calculated based on the corresponding CVSS factors value. As shown in Fig. 5, we have specified the respective factor for the vulnerability ‘Violation of customer private information’ and set values for each factor of the three metric groups of CVSS. We note that this vulnerability affects a lot (AC: 0.35) the Network (AT: 1). It has no impact in Confidentiality (C: 0), but affects completely Integrity (I: 0.660) and partially Availability (A: 0.275). It keeps the same values as the previous vulnerability at the temporal factors. At the environmental factor, we notice that this vulnerability does not damage system (CDP: 0) and has a medium distribution (TD: 0.75). It affects security requirements by augmenting integrity completely and availability partially and has no impact in confidentiality ([CR, IR, AR], [0.5/1.5/1]). With these values, we will calculate the Base, Temporal and Environmental score.

BaseScore = 5,89
 Temporal Score = 5,27
 Environmental Score = 3,87

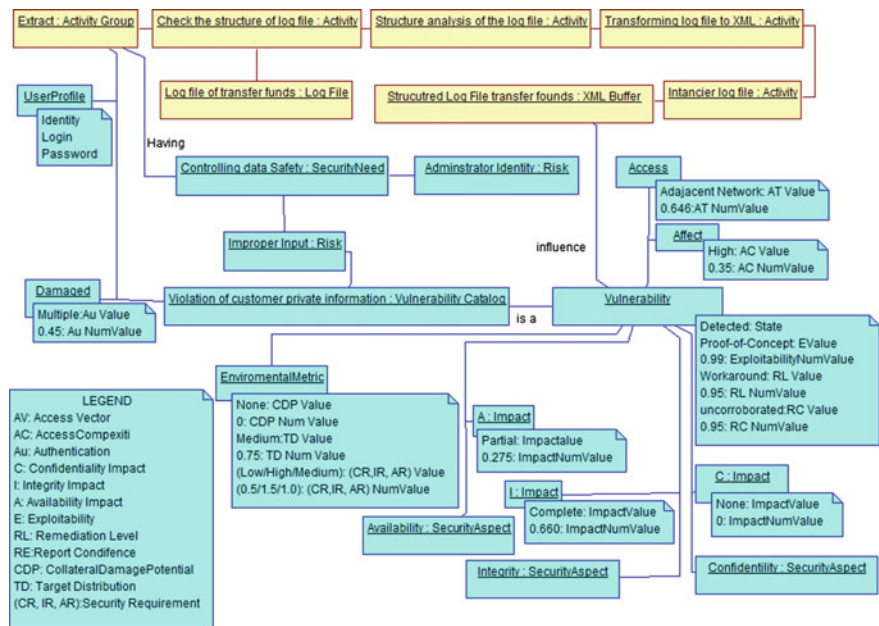


Fig. 5 ‘Violation of customer private information’ security factors in the “Log File Structure” activity diagram

Knowing that the base metric group is the mandatory score in CVSS, we compare the two calculated results. The base score of Violation customer private information is greater than the base score of ‘Unprotected transport credentials’. This observation leads us to consider that the first vulnerability is more severe and it impact highly the process. It should be treated quickly and strict and relevant security solutions should be proposed to avoid any attempt to attack private information of bank’s customer.

4.2 Transformation Process

In ETL design process, the transformation step is an important and complex task. There are a set of web transformation activities specific to web sources. These activities take as input web data sources structured in the Extraction process. At this level, the data should be protected from intrusion and undesirable access. These risks affect the data quality and safety. In Table 2, we define and categorize the anticipated vulnerability of the transformation process.

Table 2 Security risks in the WEB transformation process

Risk	Vulnerability
Network problem	Lack of access control between process
	Risk of loss data
	Problem of downtime during the treatment and/or the transfer of data
	Data corruption during the transmission to the Loading process
	Risk of losing business data (Customer Private data, Accounts number,...)
Data safety	Joining erroneous data with incorrect values or misspelled attribute
	Missing encryption of accounts information
	Missing encryption of customer private information
	Improper autorisation
	Transmission of sensitive data to the loading process without encryption
	Erroneous business web data
	Missing encryption of sensitive data
	Cybercrime: Phishing

This table represents a catalog helping the security manager to find the anticipated vulnerabilities and to define for each one the corresponding security remediation. This goal is achieved by relying the CVSS factors value prioritizing vulnerability and facilitating the choice of the security solution.

4.3 Load Process

Finally, a secure environment should be available to the loading process which loads the processed data to the WeBhouse. In this stage, many problems are detected and should be remediated to assure the WeBhouse data safety. In Table 3, we list some security problems which can be detected in the loading process.

Table 3 Security risks in the WEB Loading process

Risk	Vulnerability
Data safety	Risk of loss data
	Downtime problem during the data loading
	Unprotected DSA

5 Conclusion

An early detection of security requirement has an impact on the further design decisions providing better security specifications and final products. Security constraints should be considered in the whole development process from early stages to final tools. Because the ETL processes present the first stage in the WebHouse development and it treats an important volume of data coming from heterogeneous and uncontrolled sources, our challenge is to secure the WEB ETL processes and offer a safe environment to modulate the extracted data. In this paper, we propose an approach to secure Web ETL processes by describing the anticipated vulnerabilities for each step and measuring the Score of each one based on the CVSS factor. This measuring prioritizes the critical vulnerabilities and helps the security manager to define the suitable solutions. Our future work consists in the implementing of our proposal and defining a tool helping designer from meta to model while treating non-functional security requirements from the business level.

References

1. OMG, O.M.: Omg unified modeling language (omg uml), superstructure, v2.1.2 (2007)
2. Mell, P., Scarfone, K., Romanosky, S.: Common vulnerability scoring system version 2.0, NIST and Carnegie Mellon University, 1st edn, June 2007
3. Mehedintu, A., Bulgiu, I., Pirvy, C.: Web-enabled Data Warehouse and Data Webhouse
4. Hernandez, P., Garrigs, J.: Model-driven development of multidimensional models from web log files, ER'10 Proceedings of the international conference on Advances in conceptual modeling: applications and challenges, pp. 170–179 (2010)
5. Liu, J., Hu Chaou, J., YuanHeJin: Application of Web Services on The Real-time Data Warehouse Technology (2010)
6. Kimball, R., Merz, R.: Le DATA WEBHOUSE:Analyser les comportements client sur le Web, Eyrolles Edition, 2000
7. Muralini, M., Kumar, T.V.S., Kanth, K.R.: Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes, In Third UKSim European Symposium on Computer Modeling and Simulation, pp. 142–147 (2009)
8. Muralini, M., Kumar, T.V.S., Kanth, K.R.: Simulating: Secure ETL Process Model: An Assessment of Security in Different Phases of ETL, In Software Engineering Competence Center (2013)
9. Kiran, P., Sathish Kumar, S., Kavya, NP.: Modelling Extraction Transformation Load embedding Privacy Preservation using UML, Int. j. comput. Appl. (2012)
10. National Institute of Standards and Technology Special Publication 800–30, Risk Management Guide for Information Technology Systems, June 2001
11. National Institute of Standards and Technology Special Publication 800–53, Recommended Security Controls for Federal Information Systems, December 2007
12. National Institute of Standards and Technology Special Publication 800–55, Performance Measurement Guide for Information Security, July 2008
13. Cheng, P., Wang, L., Jajodia, S., Singhal, A.: Aggregating CVSS Base Scores for Semantics-Rich Network Security Metrics, In SRDS, 2012, pp. 31–40
14. Pengsu, C., Lingyu, W., Anoop, J.: Aggregating CVSS Base Scores for Semantics-Rich Network Security Metrics, pp. 31–40. IEEE, SRDS (2012)

15. Siv, H. Virginia, H., Franqueira, N.L., Erlend A. Engum.: Quantifying security risk level from CVSS estimates of frequency and impact, *J. sys. softw.* 83 (9), ISSN 0164-1212, pp. 1622–1634 (2010)
16. Mallek, H., Walha, A., Faiza, G.J., Gargouri, Faiez: ETL-Web process modeling, 8me edition de la confrence sur les Avancs des Systmes Dcisionnels, Hamamet Tunisia (2014)
17. Bellovin, S.: On the Brittleness of Software and the Infeasibility of Security Metrics, *IEEE Security and Privacy* (2006)
18. Thompson Lord Kelvin, W.: Electrical Units of Measurement,? Lecture at the Institution of Civil Engineers, London, 3 May 1883, *Popular Lectures and Addresses*, vol. 1, pp. 73–136 (1889)



<http://www.springer.com/978-3-319-23466-3>

Computer and Information Science 2015

Lee, R. (Ed.)

2016, XIII, 252 p., Hardcover

ISBN: 978-3-319-23466-3