SIMULATING SECURE DATA EXTRACTION IN EXTRACTION TRANSFORMATION LOADING (ETL) PROCESSES

Ashish Kumar Rastogi

Department of Information Technology, Azad Group of Technology & Management.

Lucknow (India)

ABSTRACT

Extraction-Transformation-Loading (ETL) tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse. In this paper, we propose a methodology for the earliest stages of the data warehouse design, with the goal of tracing the analysis of the structure and content of the existing data sources and their intentional mapping to the common conceptual data warehouse model. The methodology comprises a set of steps that can be summarized as follows: (a) identification of the proper data stores; (b) candidates and active candidates for the involved data stores; (c) attribute mapping between the providers and the consumers, and (d) annotation of the diagram with runtime constraints.

I INTRODUCTION

ETL is an abbreviation of the three words Extract, Transform and Load**.** It is an ETL process to extract data, mostly from different types of systems, transform it into a structure that's more appropriate for reporting and analysis and finally load it into the database and or cube(s).

ETL – Extract from source, in this step we extract data from different internal and external sources, structured and/or unstructured. Plain queries are sent to the source systems, using native connections, message queuing, ODBC or OLE-DB middleware. The data will be put in a so-called Staging Area (SA), usually with the same structure as the source. In some cases we want only the data that is new or has been changed, the queries will only return the changes. Some ETL tools can do this automatically, providing a changed data capture (CDC) mechanism.

ETL – Transform the data, once the data is available in the Staging Area, it is all on one platform and one database. So we can easily join and union tables, filter and sort the data using specific attributes, pivot to another structure and make business calculations. In this step of the ETL process, we can check on data quality and cleans the data if necessary. After having all the data prepared, we can choose to implement slowly changing dimensions. In that case we want to keep track in our analysis and reports when attributes changes over time, for example a customer moves from one region to another.

ETL – Load into the data warehouse, finally data is loaded into a data warehouse, usually into fact and dimension tables. From there the data can be combined, aggregated and loaded into datamarts or cubes as is deemed necessary.

In the information technology profession, we have seen few of the applications like order processing, general ledger, inventory, in-patient billing, checking accounts, insurance claims, and so on. These applications are very popular systems which run businesses. They process orders, maintain inventory, keep the accounting books, service the clients, receive payments, and process claims. Without these computer systems it is very difficult to survive in today's modern generation.

As businesses grew more complex and complex, corporations spread gradually. The operational computer systems did provide information to run day-to day operations, but what the executives needed were different kinds of information that could be readily used to make strategic decisions. The operational systems, is important as they were not able to provide strategic information. Data warehousing is a new paradigm specifically intended to provide vital strategic information. According to the author Bill Inmon the term "Data Warehouse" is defined as follows

"A data warehouse is a subject-oriented, integrated, time-variant and non volatile collection of data in support of management's decision making process".

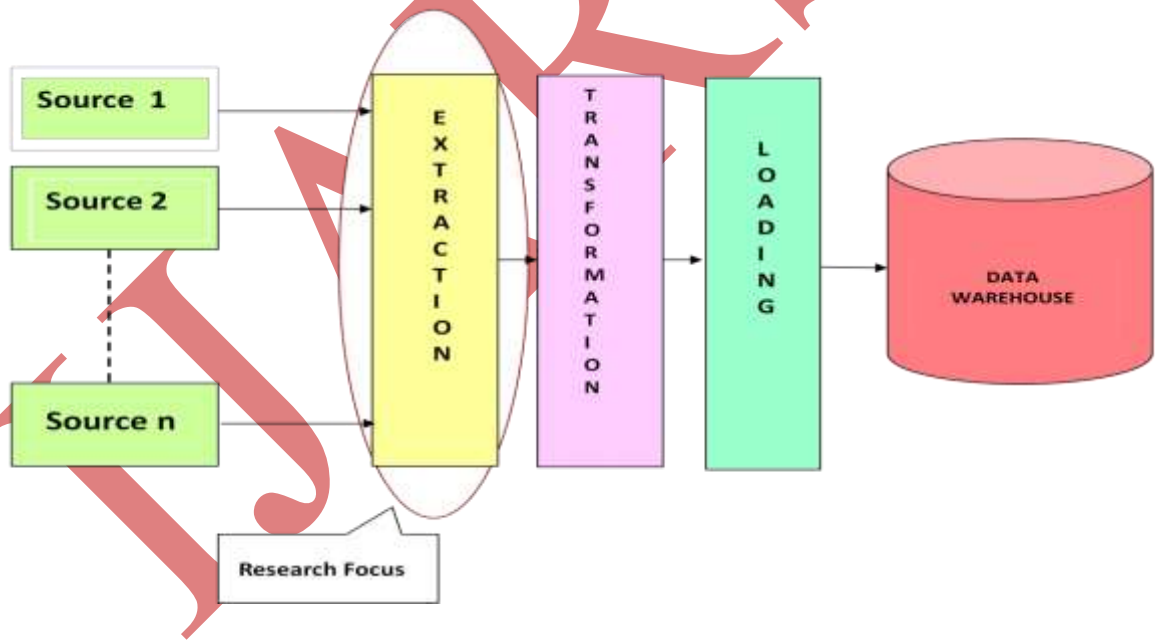

Figure 1 ETL Architecture

"Subject-oriented" here means that the data addresses a specific subject such as sales, inventory, etc."Inventory" means that the data is obtained from heterogeneous sources. "Time variant" implies that the data is stored in such a way that when some data is changed, then that data has been changed is also stored. "Non-volatile" implies that data is never removed, i.e., historical data is also kept. So, whenever a change takes place in any of the field values of a table, the previous values also need to be kept in addition to the present values.

The Data warehouse obtains the data from various data sources. The data from these sources are then converted into a form suitable for data warehouse. This process is called *Extraction Transformation and Loading (ETL)* of data into target database.

II DATA EXTRACTION WITH VARIABLE DATASETS

Data warehouse projects involve populating databases with unique subsets of records. These record subsets are frequently extracted from much larger sets that can sometimes contain many millions of records. Data extractions is based on lookup tables, which can be implemented using several different techniques and their performance differences are examined using several test data sets and flat files. This can help to guide the warehouse designer's decisions regarding sorting, indexing, and choice of an extraction method.
Observation shows total times to extract 10,000 records from source data sets and files of 10,000, 100,000 and 1,000,000 records on the PC.

2.1 Limitation

Data for the largest merge extraction is not shown because it need high specification computer to able to sort or index the source data set with one million records. This does make the point that these operations can be very resource intensive and will become impractical at some point on any computer.

Extractions with an indexed key take the greatest amount of time. The difference is substantial, with key times of 20 minutes to extract 1,000,000 records, versus one or two minutes for the other techniques. The SQL extraction is the fastest technique, in part because it avoids sorting or indexing overhead.

III SECURITY CHALLENGES IN DATA EXTRACTION

Data warehouse poses its own set of challenges. One major challenge, Organization data warehouse is very large system serving different security need for different users. Data warehouse thus need some security from unauthorized access. Security most popularly is introduced, once the data warehouse is created. But the data which enters the data warehouse is also prone to violate the security needs, as it could be hacked during the building process of Data warehouse. In order to overcome this kind of security violation, certain measures should be taken during the building process of data warehouse, so that the data, which enters the data warehouse, is not hacked during the process of Extraction or Transformation or Loading.

This research work proposes the security measures during the first phase of the building process of data warehouse. This process refers to as Data extraction process.

IV SECURITY NECESSARY FOR A DATA WAREHOUSE

Many of the basic requirements for security are well-known, and apply equally to a data warehouse as they would to any other system: The applications must prevent unauthorized users from accessing or modifying data; the applications and underlying data must not be susceptible to data theft by hackers; the data must be available to the right users at the right time; and the system must keep a record of activities performed by its users. These requirements are perhaps even more important in a data warehouse because a warehouse contains

data consolidated from multiple sources, and therefore, from the perspective of an individual trying to steal information, a data warehouse can be one of the most lucrative targets in an enterprise. In addition, a robust security infrastructure can often vastly improve the effectiveness or reduce the costs of a data warehouse environment.

Some typical customer scenarios for data warehouse security include the following:

- An enterprise is managing a data warehouse that will be widely used by many divisions and subsidiaries. This enterprise needs a security infrastructure that ensures the employees of each division be able to view only the data that is relevant to their own division, while also allowing employees in its corporate offices to view data for all divisions and subsidiaries.
- An enterprise's data warehouse stores personal information. Privacy laws may govern the use of such personal information. The data warehouse must handle data so as to adhere to these laws.
- An enterprise sells data from a data warehouse to its clients. Those clients may view only the data to which they have purchased or to which they have subscribed. They should never be permitted to see the data of other clients.

The researcher say that the security is a crucial aspect for any modern software system, to ensure security in the final product, security requirements should be considered in the entire software development process. How to integrate security requirements into the analysis phase. The software security problem has been addressed by numerous approaches in the requirements phase including scenario-driven requirements analysis misuse case driven elicitation of non-functional requirements.

There are many ways to apply the security to the file in order to protect the sensitive data in the file. Some of the existing methods to protect the file during the transfer of the file from source to destination are as discussed below.

4.1 Encrypt and Decrypt Text Files Using Ultra Edit

Ultra Edit's built-in encryption provides a quick and easy way to encrypt/decrypt the sensitive data , allowing us to keep our sensitive data secure. As of v14.00, we can encrypt and decrypt our files using a built-in advanced encryption method.

4.2 Structured Binary Encryption

The flat file can be encrypted as a binary file and forwarded to the destination. The destination uses this binary file and will convert back to flat file only when the destination is aware of the database structure. Only the structure of the file is used to decrypt the file back to flat file.
Even if the hacker hack the file during the extraction process he will fail to understand the meaning of the data as he will not be aware of the database structure. The database structure is transferred to the target system only once and the details of the database structure would be few bytes thus is quiet efficient to transfer the same to the target system.

As mentioned in the chapter 3, Fixed Width Flat Files are having fix field length.

Example

Struct EMPLOYEE

{

ID; // maximum field length 04 digits

DoB[10]; // fixed field length 10 characters like 14/12/1977

Name [100]; // maximum field length 100 charcters

Division [35]; // maximum field length 04 characters

Address [300]; // // maximum field length 100 charcters

}

The database structure for the first row of the flat file with field size as mentioned below separated by delimiters is sent to the destination separately only once.

Table 1: ASCII to Binary Conversion Table

Text	ASCII	Binary
K	chr(75)	01001011
L	chr(76)	01001100
M	chr(77)	01001101
N	chr(78)	01001110
O	chr(79)	01001111
P	chr(80)	01010000
Q	chr(81)	01010001
R	chr(82)	01010010
S	chr(83)	01010011
T	chr(84)	01010100
U	chr(85)	01010101
V	chr(86)	01010110
W	chr(87)	01010111
X	chr(88)	01011000
Y	chr(89)	01011001
Z	chr(90)	01011010
[	chr(91)	01011011
\	chr(92)	01011100
]	chr(93)	01011101
^	chr(94)	01011110
_	chr(95)	01011111

4.3 Flat File Checker (Flafi) For Data Validation

Flat File Checker (FlaFi) is a simple and intuitive tool for validation of structured data in flat files (*.txt, *.csv, etc.). It is the best application to change the way you validate data and make processes easy and efficient. FlaFi's intuitive interface makes it easy to translate business rules into Flat File Schema that defines validation criteria. You don't need any product specific knowledge to start using the application, which is often the case with ETL software.

FlaFi is essential to ensure painless data exchange without errors. It minimizes effort

- Data quality with virtually no manual involvement
- Interface specification for data exchange created on-the-go
- Successful data exchanges with no repetitive interactions with the external data providers (as long as they

screen the data against your Schema prior to supply) and saves time.

- Minimum turnaround time due to guaranteed quality of each data supply
- Faster set-up of new exchange processes

V ALGORITHM DESIGNED FOR DATA EXTRACTION

1. List all the source systems connected to the destination machine is popped at the target machine.
2. The target machine selects the source system from the above list and check for its availability by pinging to that client.
3. Once the selected client is available connect to that client.
4. Select from the client, which data base file to be extracted.
5. After selecting the file to be extracted, convert them to flat file such as Spreadsheet files or word file. So on…
6. Apply the pass code to the file as a security measure in order to protect the file from hacking during the transfer of file from the source to the target machine.
7. Now extract the file to the target machine.
8. After the file is received at the target machine, the user at the target need to apply the same pass code to open the flat file and convert back to the database for further process.

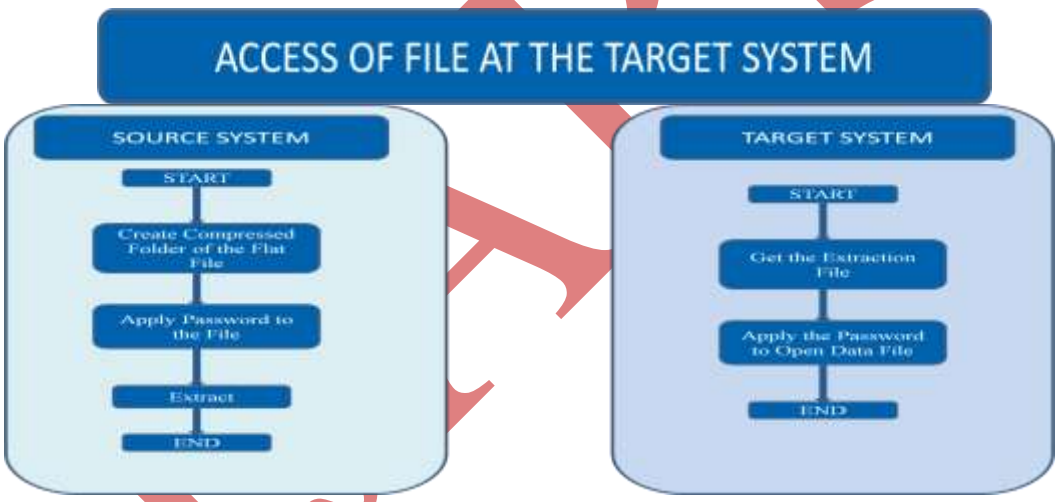

Figure 2: Access of Flat File at the Target System

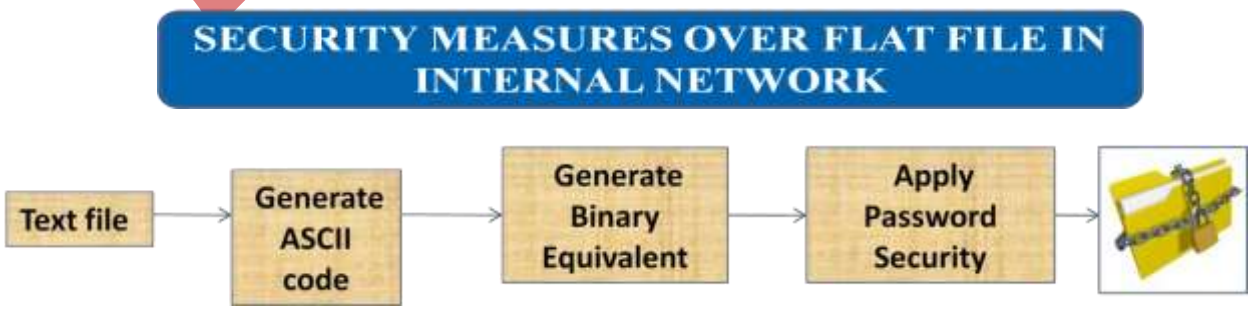

Figure 3: Security Measure in External Network, Structured Binary Encoding

The target system first looks for the valid source systems from which the data to be extracted. Once the list is ready at the target system, it will select the source from which the data is to be extracted and check whether the source is currently available for extraction by executing a ping command from the target system. Once the source is available for the service the target system is connected to the source system.

Target system selects the file which has to be extracted from the source and converts the file to flat file. Apply the pass code to the file and the file can be extracted to the target system. During the transfer of file from source to target system, the hacker cannot access the file as it is pass code protected.

At the target system, the same pass code is applied to open the flat file and then the flat file is converted back to normal data base file for further procedure. If the file is extracted for the first time from the source machine then full extract method is used to extract from the source to the target machine. Full extraction is an extraction method where in the entire file is extracted from the source.

If the updates are done over a file which is earlier extracted by the target machine then only the updates are to be extracted from the source to the destination machine. This is done using change data capture method, where only the updates which have occurred over a source database are to be extracted to the target machine. This is implemented using triggers which are raised by the source to the destination machine. After the trigger is raised at the destination, the target machine will select the source from which the trigger had been raised and will check for the availability of the machine and then connect to the machine, if available. After the successful connection, all the updated records are extracted from the database in the form of a flat file and extracted in a similar fashion as explained earlier.

The updates are reflected even at the target system, after the update records are extracted from the source system.

VI CONCLUSION

Security to the data is one of the major challenges and area of concern in today's world. Current approaches for the modelling of ETL do not address the security issues in the ETL modelling. This research work proves and shows the improvement in data extraction speed by using flat files with security measures.

The extraction time for smaller number of records such as 100 records does not show much difference in the extraction time for database file or flat file. The extraction speed for Database file consisting of 100 records takes 1ms, whereas the same 100 records can be extracted in .99 ms using flat files which are not much difference in the extraction speed.

When number of records is increased above 2000 records the difference in extraction time using flat file makes huge difference and makes sense in improving of the extraction process of ETL rather than using extraction by database file. The extraction time for 2000 records using flat files takes 16.37 ms whereas the extraction time for the same takes 25ms using database file.

Thus the improvement in the extraction process with respect to space and time domain is achieved successfully during this research work.

Current approaches for the conceptual modelling of ETL do not address the security aspects in the conceptual modelling phase. This research work proposes security to the data during the building process of the

warehouse. The building process constitute of ETL i.e. Extraction, Transformation and Loading. This research proposes security and improvement in the first phase of extraction process of ETL. Protection to the file to be extracted from the source is implemented by applying pass code to the file. Once the file is extracted at the destination the same pass code is applied to open the file at the target machine. The user needs to remember the password which he applied before extraction, for him to open the file after extraction by applying the same pass code at the target machine.

Improvement during extraction process is achieved by converting the data base table into a flat file and extract. The flat file needs less storage space on the disk compared to data base table. The extraction using flat file is much faster than extraction of the database table.

The extraction mechanism used is full extraction if the database is extracted for the first time and change data capture method, if the database is the modified once.

We have observed that the data extraction using direct database file, requires more time compared to that of the flat file database. This research method of extraction does provide security to the flat file, thus the hacker cannot make use of the content of the file during the extraction process.

This work can be extended further for the later building process of the data warehouse for transformation and loading process of the ETL. Data for the largest merge extraction is not shown because it need high specification computer to able to sort or index the source data set with one million records. This does make the point that these operations can be very resource intensive and will become impractical at some point on any computer.

REFERENCES

[1] Javed, M.Y.; Nawaz, A.; Second International Conference on Data Load Distribution by Semi Real Time Data Warehouse Computer and Network Technology (ICCNT), 2010 Digital .

[2] Ying Pei; Jungang Xu; Qiang Wang; 2nd International Workshop on One CWM-Based Data Transformation Method in ETL Process Database Technology and Applications (DBTA).

[3] 2nd International Workshop on ETL Function Realization of Data Warehouse System Based on SSIS Platform Database Technology and Applications (DBTA).

[4] Xi-Qian Chen; Zhong-Xian Chi; Xiu-Kun Cao; International Conference on Applying DP to ETL of spatial data warehouse Machine Learning and Cybernetics, 2004.

[5] Maddodi, S.; Attigeri, G.V.; Karunakar, A.K.; 3rd International Conference on Data De- duplication Techniques and Analysis Emerging Trends in Engineering and Technology (ICETET).

[6] Jian-hua Luo; Yong-ming Chen; Qing-ling Zeng; International Conference on The Design and Implementation of Electric Power Data Integration System Based on the Extraction- Transformation-Loading Technology Management and Service Science (MASS).