Empirical Models for the Performance of ETL Processes

M Mrunalini T V Suresh Kumar K Rajani Kanth
Department of Computer Applications
M S Ramaiah Institute of Technology
Bangalore, India

ABSTRACT
Generally, software projects' outcomes will give us various aspects of quality parameters. In such cases, empirical studies with prototyping exercises are well suited to analyze/understand the system. ETL (Extraction-Transformation-Loading) is the software responsible for extracting data, cleaning, transforming and loading the data into a data warehouse. ETL is a large software system. The performance of the decision support system depends on the data warehouse that it uses. ETL tools play a major role in building the data warehouse; these tools need to have good performance in order to improve the performance of the whole system. An experimental study is conducted to analyze the performance of the ETL tool. Two ETL tools are considered; one with integrated security and another without integrated security. The time for data extraction in different environments is recorded. Further, regression analysis is done on the experimental data and observed the behavior of the tools and developed the empirical models. Both tools have shown the same behavior in performance for different extraction data sizes.

General Terms
Empirical Models, Performance

Keywords
Regression Analysis, Integrated Security, Secure ETL, Performance Analysis, Experimental Study.

1. INTRODUCTION
Data warehouses are complex systems whose main objective is to facilitate the information for decision making process of Business Intelligence (BI) workers. Decision support systems are becoming increasingly more critical to daily operations of organizations. The complexities of decisions in the information age compel every decision maker to utilize large information available for supporting business decisions. Data warehouses have become the focal point for decision support in organizations today. ETL (Extraction-Transformation-Loading) processes are responsible for the extraction of data from heterogeneous operational data sources, their transformation including homogenization, cleaning, normalization etc. and loading into data warehouses [1].

A data warehouse contains the sensitive information, which is used for decision making process, it is necessary to take precautionary measures in the data warehouse building process where ETL process plays major role [2]. At the same time heterogeneity and performance issues are major challenges.

ETL processes must be completed in a certain time window; thus, it is necessary to optimize their execution time [3].

The key challenge to build a data warehouse involves the ability to turn data into easily and quickly accessible information. This challenge is precedent by the tasks of extracting, cleaning, transforming data into usable repositories. [4]

There are two approaches of building the data warehouse. (i) Top-down approach and (ii) Bottom-up approach [5]. The second approach i.e., bottom–up approach, the data warehouse is built based on the query posed by the user. Especially in this case, the ETL tool should be fast enough to extract the required data and build the data warehouse. The problem of low performance in data extraction can be critical in the BI projects. If a BI report is taking a lot of time to run or the data displayed are no longer available for taking critical decisions, the project can be compromised [6].

While populating the data warehouse, apart from heterogeneity and performance issues security is also an important non-functional requirement that need to be addressed. Hence, several aspects of information security on ETL processes are to be realized, for example Confidentiality, Integrity, Inference, Authentication, Data Corruption etc.

In general security is considered to be an afterthought of a software product. However, the quality parameters like response time and throughput require considerable attention during implementation of ETL. Keeping this in view there should be a tradeoff between security and performance of the software that has to be studied during implementation of ETL. Since ETL is a complex system and it is required to build quality information for data warehouse, it is imperative to consider an integrated approach addressing security of ETL and intern assessing performance. Current literature shows that the available models to build data warehouse do not adequately address the security modeling. The issues in security modeling include, devising a reliable and flexible security model that widely considers all the components of the warehousing architecture, including ETL and data sources [7]. It infers that there is a need for integrated security model for ETL process. When security is added as a functional requirement, the size and complexity of the software will also increase with varying performance.

The main objective of empirical software engineering research is to transfer the research results to industrial practices. Two main problems in this transfer are the lack of explanatory power and the lack of realism of controlled experiments. It is difficult to increase the explanatory power of case studies, where as there is large scope for increasing the realism of controlled experiments. To convince industry about the validity and applicability of the experimental results, the experimental tasks, elements and the execution environments of the experiments should be as realistic as practically possible [8].

In this paper, two ETL tools are considered; one with integrated security and the other without integrated security. Experimental studies are conducted in two software companies and recorded the time for data extraction for different data sizes. Data extraction time is considered as

response time of extraction phase of the ETL tool. Using the experimental data, empirical models are developed to analyze the performance of the ETL tools at extraction phase. In Section 2, a brief study on the available literature is presented. In Section 3, the experimental study is presented. In Section 4, the results of the study are discussed and in Section 5 summary of the paper is presented.

2. LITERATURE SURVEY

The open issues related to the modeling and design of data warehouses are discussed in [7]. On the basis of the outcome of the workshop and survey, the authors of [7] conclude that though modeling and design of data warehouses have been investigated for about a decade, several important challenges still arise. Moreover, the outcome of the data warehouse workshop [7] says that, due to the strategic importance of data warehouses, it is absolutely crucial to guarantee their quality from the early stages of a project. While some relevant work on the quality of data has been carried out (e.g., [9, 10]), there is still no general agreement on the quality of the design process and its impact on decision making.

The authors in [11, 12] present the modeling and optimization of ETL processes at the logical level. A design method that includes an algorithmic transformation of conceptual to logical models for ETL processes is discussed in [13]. In [14], the authors give a conceptual model of the ETL processes. In [15, 16] the authors focus on the dynamic [15] and static [16] modeling of the ETL processes. An UML based approach for the modeling of ETL processes is discussed in [17]. The problem of determining the best possible physical implementation of an ETL workflow, given its logical level description and an appropriate cost model as inputs, is discussed in [18]. The generic frame work on conceptual and logical models of ETL is discussed in few more papers [19, 20, 21, 22].

It is observed from the literature that assessment of security and performance are not adequately addressed in the context of ETL and this needs to be addressed. Due to the strategic importance of data warehouses, it is absolutely crucial to address security and performance together. The current research and development on data warehouse/ETL do not adequately address the impact of security on the performance. Hence it is required to study the impact of security on the performance of the ETL process in detail.

3. EXPERIMENTAL STUDY

3.1. System Model

In general, the architecture of data extraction in ETL processes is shown in Figure 1. The components Integrator and Wrapper play a major role in the data extraction process. Wrapper establishes the connection with the data sources and extracts the data from the source and informs the Integrator. Integrator stores the extracted data in the Data Staging Area (DSA).

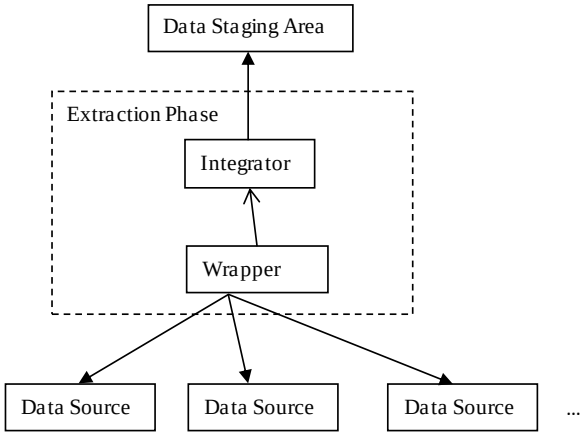

Fig 1: Architecture of Data Extraction in ETL Processes

Further to assess the process suitable deployment environment is required. The deployment environment of the experiments is shown in Figure 2. The experiment is conducted in Windows and Linux environments. All the servers are connected with 1000 kbps LAN. They have multiple database servers like MSSQL, MYSQL, MONGO DB and ORACLE.

Performance of the ETL server (tool), shown in Figure 2, is obtained through the experimental studies.

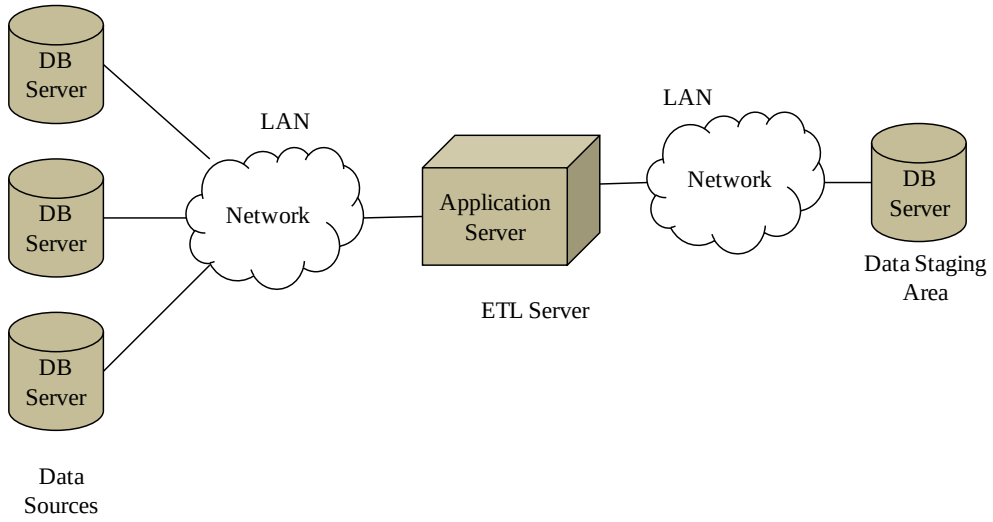

Fig 2: Architecture of the Experimental Environment

(i) Experimental Setup-1

The execution environment of the secure ETL prototype (ETL tool-1) is as follows.

Platform: Windows

Client Configuration: Intel(R) Core(TM) 2 Duo CPU E6550 @ 2.33GHz, 4 GB RAM

Server Configuration: Two Intel Xeon Dual Core 3.00 GHz, RAM 8 GB, Hard disk 500 GB

Connection Environment: LAN

Source DB: SQL Sever

Destination DB: SQL Sever

(ii) Experimental Setup-2

The deployment environment of the ETL tool-2 (ETL without integrated security) is as follows.

Platform: Linux

Client Configuration: Pentium(R) Dual-Core CPU, 4 GB RAM, 500 GB HD @ 54000rpm

Server Configuration: Intel Core i7 processor, 8 GB RAM, 500 GB HD @ 54000rpm

Connection Environment: LAN

Source DB: MySql

Destination DB: MongoDB (single instance)

3.2. Empirical Models

Empirical studies are a key way to get information about quality parameters and move towards well-founded decisions. Empirical studies take many forms. They are realized not only as formal experiments, but also as case studies, surveys, and prototyping exercises as well [23].

Performance of the ETL tool using empirical study at extraction phase is analyzed using the following procedure.

Algorithm

Step 1: Connect to the data source

Step 2: Select the data to extract

Step 3: Extract the data and store in Data Staging Area

Step 4: Record the extraction time and extracted data size

Step 5: Analyze the data

5.1 Analyze the distribution of data set

5.2 Apply suitable statistical test for goodness of fit

5.3 If the considered distribution is proved go to Step6

Otherwise find the alternate distribution and repeat Step 5

Step 6: Build a relation (empirical model) for the data

Step 7: Estimate the performance of the ETL tool

A prototype tool (ETL tool-1) for secure data extraction is developed and using this prototype the experimental study-1 in a software development company has been carried out to observe the performance behavior of the secure ETL tool. During the experiment, extraction time for different sizes of data is recorded and the results are presented in Table 1.

Table 1. Data Extraction Time of Secure ETL

Sl. No	Extracted Data Size (in KB)	Extraction Time (in Seconds)
1.	8	0
2.	16	0.22
3.	272	5
4.	816	21
5.	1264	84
6.	2640	374
7.	3528	474
8.	66232	2033
9.	92678	4832
10.	154896	7866

From Table 1, it can be observed that the change in extraction time as the extracted data size increases. The table shows that even when there is drastic increase (exponential) in the extracted data size, the extraction time increases slowly (not at the pace of data size variation). Further it infers that the performance of the tool varies at different data sizes and can term as clusters of similar behavior. This can be observed more clearly from Figures 3 and 4.

To test whether the data size is exponentially distributed or not, Kolmogorov-Smirnov test is used for exponential distribution. Since the performance of the tool is changing at different data sizes we divided the data into two subsets (Reading 1 to 5 as Set 1 and readings 6 to 10 as Set 2) and tested the distribution.

The null hypothesis is formed as follows:

H_0: the data sizes are exponentially distributed.

If the data is exponentially distributed then the critical value $D_{n,\alpha}$ will be larger than D_n. From the Kolmogorov-Smirnov table it's seen that for Set1 $D_{n,\alpha} = D_{5,0.05} = 0.565$ and $D_n=0.486$. For Set2 $D_{n,\alpha} = D_{5,0.05} = 0.565$ and $D_n=0.516$.

Since both $D_n = 0.486/0.516 < 0.565 = D_{n,\alpha}$, it is concluded that the data is a good fit with the exponential distribution.

Similarly extraction time is tested for exponential distribution using Kolmogorov-Smirnov test. For Set1 $D_{n,\alpha} = D_{5,0.05} = 0.565$ and $D_n=0.562$. For Set2 $D_{n,\alpha} = D_{5,0.05} = 0.565$ and $D_n=0.415$. Since both $D_n = 0.560/0.415 < 0.565 = D_{n,\alpha}$, we can infer that the data is a good fit with the exponential distribution.

Hence the data sizes of heterogeneous sources require clustering of data sizes for proper analysis of ETL tools.

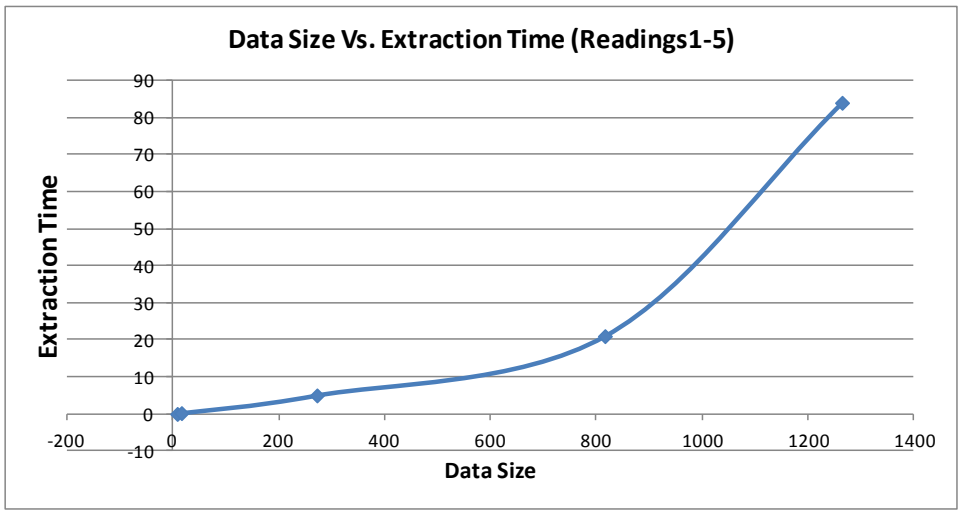

Fig 3: Data Size Vs. Extraction time for the readings 1-5 of Table 1

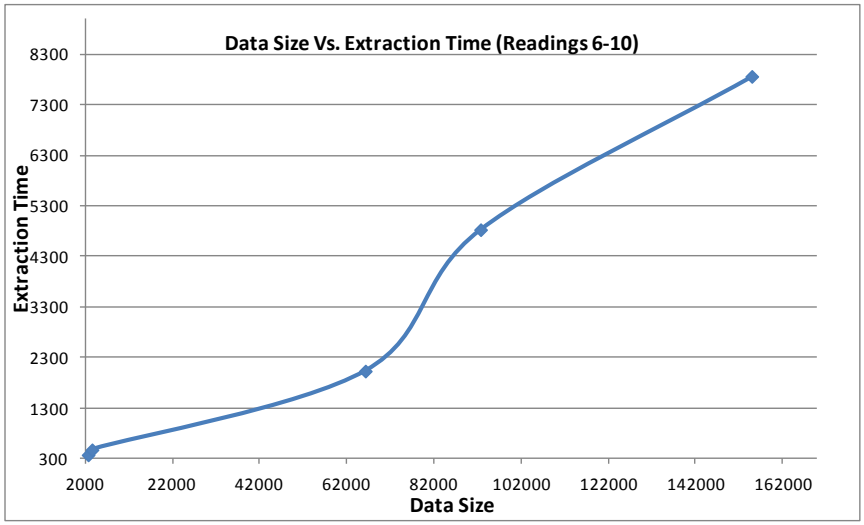

Fig 4: Data Size Vs. Extraction time for the readings 6-10 of Table 1

Further, the performance of ETL tool-1 is analyzed by performing regression analysis.

The above Figures show that, the slope of the curve changes at different levels. A regression analysis in MATLAB is carried to understand the relationship between data size and extraction time and explored the forms of these relationships.

For the experimental data in Table 1, at two relations are derived i.e., a relation for the databases with sizes less than 1 MB and another kind of relation for the databases with sizes above 1 MB.

The relations are as follows:

$Y = 0.025\,X + 1\ for\ X \leq 1MB$ ……..… (1)

with correlation coefficient = 0.9236

$Y = 0.045\,X + 30\ for\ X > 1MB$ ……..… (2)

with correlation coefficient = 0.9236

Where X is the extracted data size in KB and

Y is the extraction time in seconds

The above relations help in estimating the extraction times in the respective ranges with less than 20% of the error. Relation (2) may hold good only up to certain range. Beyond that the basic nature of change in the performance at different data sizes may require finding another relation.

The results of experimental study-2 are also considered, i.e., ETL tool (ETL tool-2 without integrated security) of another software company and observed the performance behavior of the tool. We recorded the extraction time for different sizes of data for this tool too. The results are presented in Table 2.

Table 2. Data Extraction Time of ETL (without Security)

Sl. No	Extracted Data Size (in MB)	Extraction Time (in Seconds)
1.	50	1.59
2.	104	3.04
3.	155	4.61
4.	254	7.14
5.	300	15.86

This ETL tool also shows the behavior similar to our secure ETL prototype. Table 2 shows that even when there is drastic

increase in the extracted data size, the extraction time increases slowly.

To test whether the data size of Table 2, is exponentially distributed or not, Kolmogorov-Smirnov test for exponential distribution is used.

The null hypothesis is formed as follows:

H_0: the data sizes are exponentially distributed.

If the data is exponentially distributed then the critical value $D_{n,\alpha}$ will be larger than D_n. From the Kolmogorov-Smirnov table it is seen that for $D_{n,\alpha} = D_{5,0.05} = 0.565$ and $D_n = 0.429$.

Since $D_n = 0.429 < 0.565 = D_{n,\alpha}$, it is concluded that the data is a good fit with the exponential distribution.

For extraction times the results are $D_n = 0.313$ and $D_{n,\alpha} = D_{5,0.05} = 0.565$. Since $D_n = 0.429 < 0.565 = D_{n,\alpha}$, it is concluded that this data too exponential distribution is a good fit.

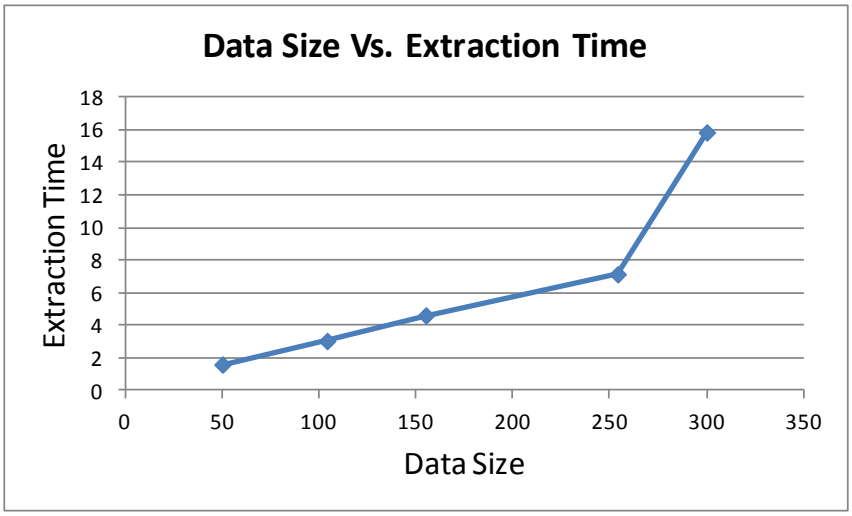

Fig 5: Data Size Vs. Extraction time for the readings of Table 2

From Figure 5, it can be observe that, performance of this tool also varies at different stages. Hence regression analysis is carried for this case too.

For the experimental data in Table 2, two relations are derived i.e., a relation for the databases with sizes less than 300 MB and another kind of relation for the databases with sizes above 300 MB.

The relations are as follows:

$Y = 0.03\,X + 0.12 \;\; for\, X < 300\,MB$ ……..… (3)

with correlation coefficient = 0.9992

$Y = 0.05\,X + 0.2 \;\; for\, X \geq 300\,MB$ ……..…… (4)

with correlation coefficient = 0.9992

Where X is the extracted data size in MB and

Y is the extraction time in seconds

The above relations help in estimating the extraction times in the respective ranges with less than 20% of the error. In this case also, Relation (4) may hold good only up to certain range. Beyond that the basic nature of change in the performance at different data sizes may require finding another relation.

3.3. Discussion

The above experimental studies show that both the tools show the same behavior. The developed prototype, ETL tool-1, with integrated security showed the similar behavior as ETL tool-2 for data extraction of different data sizes. ETL tool-1 has low performance compared to ETL tool-2 as security is added as functional requirement for ETL tool-1. It is obvious that when security is added as part of functional requirements, the size of the software increases which causes the degraded performance. However, the tradeoff analysis of security and performance is shown in the paper [24]. The results infer that though there are varying response times, ETL tools show the similar behavior in any type of execution environment. Thus it is suggested that usage of our secure ETL tool for data extraction is better as it has extra advantage i.e., security. We can analyze the performance of any ETL tool using the procedure proposed in this paper.

4. CONCLUSION AND FUTURE ENHANCEMENTS

Performance of two ETL tools is analyzed; ETL tool-1with integrated security and ETL tool-2 without integrated security at extraction phase, using empirical models that are developed based on the field data. Kolmogorov-Smirnov test is done for input data analysis. Regression analysis is also done on the experimental results and mathematical relations are developed to assess the performance of the tools at extraction phase. It can be concluded that both the tools have shown similar behavior in performance. ETL tool-1 has degraded performance compared to ETL tool-2 as ETL tool-1 is loaded with security as functional requirement. However, the tradeoff between security and performance is a different issue which is discussed in the paper [24].

Current prototype is limited to extraction phase. Further the secure ETL prototype may be extended for cleansing and transformation and the experiments may be repeated to analyze the performance of the secure ETL at transformation phase as well as assess the performance of the secure ETL tool as a whole.

5. ACKNOWLEDGMENTS

Our sincere thanks to Hasten Technologies Pvt. Ltd and Phoenix Data Sieve, who have permitted us for the experimental study. We extend our special thanks to employees of these companies, Lakshmamma, Usha, Lakshmi and Vivek who have helped in executing the tools at their company.

6. REFERENCES

[1] Ralph Kimball, "The Data Warehouse ETL Toolkit", Wiley Publications, 2006.

[2] M Mrunalini, T V Suresh Kumar, K Rajani Kanth, 2013 "Secure ETL Process Model: An Assessment of Security in Different Phases of ETL", *Proc. International Journal of Software Engineering, IJSE*, Vol. 6 No. 1, January 2013 pp 33-63.

[3] Alkis Simitsis, Panos Vassiliadis, Timos Sellis, 2005 "Optimizing ETL Processes in Data Warehouses", *Proc. of the 21st International Conference on Data Engineering (ICDE 2005)*, pp 564-575.

[4] Tony Brown, 2007 "Data Warehouse Efficiency Techniques with the SAS System", *SAS Global Forum 2007*, April 16-19, 2007 - Orlando, Florida, pp 1-10.

[5] Manole Velicanu, 2007 "Building a Data Warehouse step by step", *Informatica Economică*, nr. 2 (42)/2007, pp 83-89

[6] Ion Lungu, Manole Velicanu, Adela Bara, Vlad Diaconi¸t˘a, Iuliana Botha, 2008 "Practices for Designing and Improving Data Extraction in a Virtual Data Warehouses Project", *Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844 Vol. III, Suppl. issue: Proceedings of ICCCC 2008*, pp. 369-374.

[7] Stefano Rizzi, Alberto Abell´, Jens Lechtenb¨orger and Juan Trujillo, 2006 "Research in Data Warehouse Modeling and Design: Dead or Alive?" *Proc. DOLAP'06*, pp. 3-10.

[8] Dag I. K. Sjøberg , Bente Anda , Erik Arisholm , Tore Dybå , Magne Jørgensen , Amela Karahasanovic , Espen F. Koren , Marek Vokác, 2002 "Conducting realistic experiments in software engineering", *Proc. 1st Int. Symposium on Empirical Software Engineering*, pp 17-26.

[9] Jarke M., Jeusfeld M. A., Quix C. and Vassiliadis P, 1999 "Architecture and quality in data warehouses: An extended repository approach," *Proc. Information Systems*, vol. 24(3), pp. 229–253.

[10] Jarke M., Lenzerini M., Vassiliou Y. and Vassiliadis P, "Fundamentals of Data Warehousing," Springer Verlag, 2003.

[11] A. Simitsis, P. Vassiliadis and T. K. Sellis, 2005 "Optimizing ETL processes in data warehouses," *Proc. ICDE*, pp. 564–575.

[12] P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis and S. Skiadopoulos, 2005 "A generic and customizable framework for the design of ETL scenarios," *Proc. Information Systems*, 30 (7), pp. 492–525.

[13] A. Simitsis, 2005 "Mapping conceptual to logical models for ETL processes," *Proc. DOLAP*, pp. 67–76.

[14] P. Vassiliadis, A. Simitsis and S. Skiadopoulos, 2002 "Conceptual modeling for ETL processes," *Proc. DOLAP*, pp. 14–21.

[15] M. Bouzeghoub, F. Fabret and M. Matulovic, 1999 "Modeling data warehouse refreshment process as a workflow application," *Proc. DMDW*, pp. 6.1–6.12.

[16] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi and R. Rosati, 1998 "Information integration: Conceptual modeling and reasoning support," *Proc. CoopIS*, pp. 280–291.

[17] Juan Trujillo and Sergio Lujan-Mora, "A UML based approach for Modeling ETL Processes in Data Warehouses," in *LNCS, Springer Verlag*, vol. 2813/2003, pp. 307–320, 2003.

[18] Vasiliki Tziovara, Panos Vassiliadis and Alkis Simitsis, 2007 "Deciding the Physical Implementation of ETL Workflows," P*roc. ACM tenth international workshop on Data warehousing and OLAP (DOLAP'07)*, pp. 49–56.

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

[20] Alkis Simitsis. *Modeling and Managing ETL Processes* [Online]. Available: http://ftp.informatik.rwthaachen.de/Publications/CEUR-WS/Vol-76/simitsis.pdf

[21] Panos Vassiliadis, Alkis Simitsis and Spiros Skiadopoulos, 2002 "Logical Modeling of ETL Processes," *Proc. International Conference on Advanced Information Systems Engineering*, pp.782–786.

[22] Panos Vassiliadis, Alkis Simitsis and Spiron Skiadopoulos, 2002 "Modeling ETL Activities as Graphs," *Proc. DMDW'2002*, Toronto, Canada, pp. 52-61.

[23] Dewayne E. Perry Adam A. Porter Lawrence G. Votta, 2000 "Empirical Studies of Software Engineering: A Roadmap", *Proc. The Future of Software Engineering*, pp 345 – 355.

[24] M Mrunalini, T V Suresh Kumar, K Rajani Kanth, 2013 "Assessing the Performance and Security Trade-offs at the Early Stages of Software Development", *Proc. IndiaCom 2013*, New Delhi, India, pp 353-360.