Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004:007
DOI: https://doi.org/10.24867/21GI04Kovacevic

УПОРЕДНА АНАЛИЗА ETL АЛАТА – СИСТЕМАТСКИ ПРЕГЛЕД ЛИТЕРАТУРЕ

COMPARATIVE ANALYSIS OF ETL TOOLS – SYSTEMATIC LITERATURE REVIEW

Никша Ковачевић, *Факултет техничких наука, Нови Сад*

Област – ИНДУСТРИЈСКО ИНЖЕЊЕРСТВО И МЕНАЏМЕНТ

Кратак садржај – *ETL* (*Extraction, Transform, Load*) *алати су изузетно битни софтверски производи зато што олакшавају и подржавају ефикасно и квалитетно спровођење ETL процеса, и посредно доприносе квалитету система складишта података. У овом раду је кроз систематски преглед литературе представљено стање на тржишту у покушају да се објасни позиција комерцијалних и бесплатних алата, и да се одговори на питање правилног избора ETL алата.*

Кључне речи: *складишта података, ETL алати, ETL, пословна интелигенција*

Abstract – *ETL* (*Extraction, Transform, Load*) *tools are important software products because they facilitate and support an efficient and high-quality implementation of the ETL process, and indirectly contribute to the quality of the data warehouse system. In this paper, through a systematic review of the literature, the state of the market is presented in an attempt to explain the position of commercial and free-to-use tools, and answer the question of the correct choice of an ETL tool.*

Keywords: *data warehouse, ETL tools, ETL, business intelligence*

1. УВОД

У многим организацијама изузетно вредни подаци потпуно су неискоришћени, само зато што се налазе раштркани у различитим форматима и чувају се у међусобно неповезаним системима [1]. Због тога организације губе много новца, времена и људског напора на активности које не доносе оптималан повраћај инвестиције, а самим тим ни највећи профит. Складишта података су комплексни системи чији је основи циљ да консолидују податке из разноврсних система (извора) једне организације и евентуално њеног окружења, и да на основу тих података доносиоцима одлука пруже нове информације као основ за доношење квалитетнијих одлука.

Темељ сваког система складишта података јесте ETL (*Extraction, Transform, Load*) систем, који се користи за преузимање података из изворних система, њихову трансформацију и обраду у жељени формат и на крају за учитавање трансформисаних података у складиште података.

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је била др Соња Ристић, ред. проф.

Успешност имплементације ETL система одређује судбину сваког пројекта имплементације складишта података. Овај процес, иако није видљив крајњим корисницима, обично захтева 70% укупно потребних ресурса за имплементацију и одржавање DW (*Data Warehouse*) система [2]. Из претходне констатације се може закључити да је избор ETL алата од пресудног значаја за организације које своје пословање желе да подигну на виши ниво. У раду је извршена анализа различитих ETL алата на бази систематског прегледа литературе.

У поглављу 2 објашњен је поступак спровођења прегледа литературе и приказани су проистекли резултати. У поглављу 3 описана су ограничења прегледа, а у поглављу 4 дат је закључак рада.

2. СИСТЕМАТСКИ ПРЕГЛЕД ЛИТЕРАТУРЕ

У овом раду анализирана је постојећа литература о различитим, комерцијалним и бесплатним ETL алатима који постоје на тржишту и компарацији истих, како би се утврдила њихова тржишна позиција и стекао увид у трендове.

2.1. Планирање прегледа литературе

Циљеви систематског прегледа литературе су:

- извршити анализу различитих ETL алата и упоредити их у контексту њихове примене у креирању и одржавању складишта података;
- одабрати примарне студије које ће бити детаљно анализиране; и
- приказати резултате који ће представљати подлогу за даље истраживачке активности.

На почетку су идентификована следећа истраживачка питања (ИП):

- ИП1: Који су најпопуларнији ETL алати?
- ИП2: Да ли комерцијални алати предњаче у односу на *Open source* алате или обрнуто?

За извор података изабране су *Scopus*, *Web of science* и *Science direct* базе података, а коришћен је и *Google scholar* претраживач. Ово су најчешћи избори, с обзиром на то да садрже највећи број релевантне литературе из различитих области. Радови који су изабрани за систематски преглед литературе углавном припадају издавачким кућама као што су *Elsevier*, *Springer*, *IEEE*, али и самосталним интернационалним научним часописима.

Изабране су следеће кључне речи за формирање термина за претрагу: ETL, *Data Warehouse*, *Data Integration*, *Extract*, *Transform*, *Load*, *Survey*, *Comparison*, *Study*, *Analysis*. Сами термини за претрагу су формирани комбинацијом кључних речи и није

сваки термин укључивао сваку кључну реч. Сви термини су били искоришћени за претрагу раније наведених база података.

Основни израз за претрагу гласи: "*ETL tools*" ***AND*** "*Data Warehouse*" ***AND*** ("*Data Integration*" ***OR*** "*Extract, Transform, Load*") ***AND*** (*Survey* ***OR*** *Comparison* ***OR*** *Analysis*).

Одабрани критеријуми инклузије су:

- у радовима треба да буду представљени комерцијални или бесплатни ETL алати; и
- у радовима треба да буде упоређено најмање два ETL алата.

Критеријуми експлузије су:

- радови чији текстови нису у целини доступни;
- радови који нису написани на енглеском или српском језику;
- радови у којима се описује искључиво ETL процес;
- радови у којима је представљено потпуно ново идејно решење за имплементацију ETL алата.

2.2. Спровођење прегледа литературе

Иницијална претрага је резултирала са 117 публикација. Књиге и поглавља књига, као и чланци и текстови за које је било очигледно да су наручени од стране компанија које развијају ETL алате нису узети у обзир, да би се обезбедио већи ниво објективности.

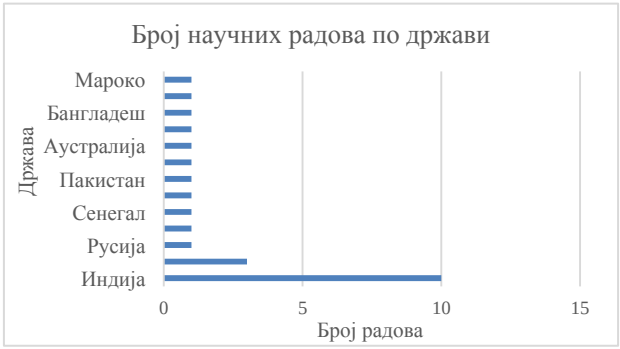
Након свих примењених критеријума остало је свега 13 радова који задовољавају услове за анализирање. Због тога су, помоћу *Google Scholar*-а, ручно претражени самостални научни часописи и доступни радови и издвојено је још 11 радова, те је коначни број анализиране литературе повећан на укупно 24 рада.

2.3. Дескриптивна статистика

Посматрајући слику 1 уочава се да не постоји изражен тренд ни у позитивном ни у негативном смеру када је у питању број објављених радова по години и посредно гледано „популарност“ ове теме. Истраживачи се са времена на време враћају на ову тему и покушавају да, у светлу нових сазнања и напретка технологије, дају комплетнији одговор.

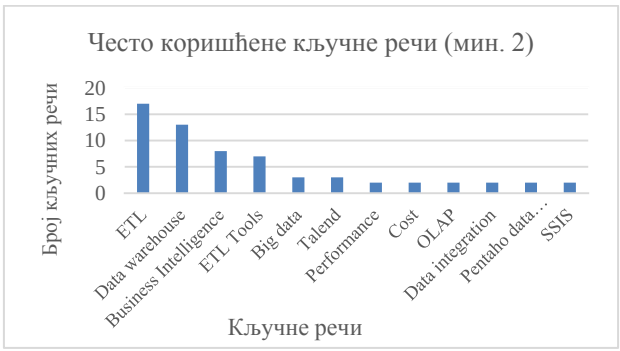
Слика 1. *Број објављених радова по години*

Државе издања представљају државе одакле долазе аутори, тачније у којима се налазе факултети и друге научне институције којима они припадају. На слици 2 је упечатљиво да Индија предњачи у односу на све остале државе, са 10 радова од 24. Сједињене Америчке Државе су друге са 3 рада и све остале државе доприносе са само једним радом. У односу на континент, највише доприноса даје Азија затим Северна Америка и Европа, али има и радова из Африке и Аустралије.



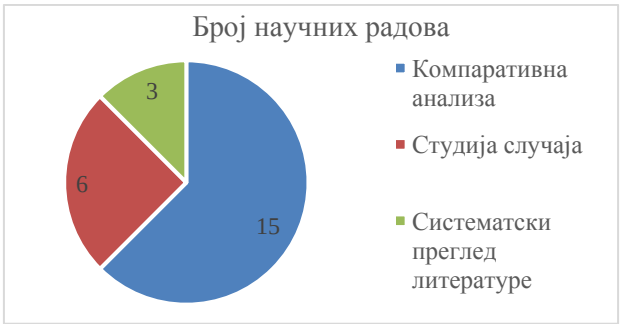
Слика 2. *Расподела радова по држави*

Извучене су све кључне речи које су наведене у примарним студијама, а на слици 3 се могу видети оне које су се нашле у најмање два рада. Као што се може претпоставити, “ETL” и „*Data Warehouse*” су најчешће (71% и 54% заступљености у радовима, редом). „*Business Intelligence*” (33%) и „ETL *tools*” (29%) су такође присутне, а још неке кључне речи су „*Performance*”, „*Cost*” и „*Data Integration*”.



Слика 3. *Најчешће коришћене кључне речи*

На слици 4 приказана је дистрибуција примарних студија у односу на истраживачке методе које су њихови аутори користили. У највећем броју случајева коришћена је компаративна анализа, док је свега четвртина радова обухватила и студију случаја, где су аутори заиста пробали да имплементирају два или више алата и упореде их на тај начин у свом истраживању.



Слика 4. *Дистрибуција радова у односу на истраживачке методе*

У будућности би требало више радити на компарацији алата у реалном окружењу, на конкретним примерима. Није пронађен ниједан рад у ком је спроведена анкета како би се, између осталог, увидело који ETL алати се користе у организацијама одређене индустрије на неком тржишту. Информације

које би се могле прикупити испитивањем релевантних доменских експерата сигурно би биле врло значајне и могле би да пруже нови, искуствено подржани поглед на ову тему.

2.4. Обухваћени ETL алати

У оквиру систематског прегледа литературе анализирани су следећи радови [1, 3–25]. Аутори изабраних радова су поредили најмање два алата, али већина њих је поредила и доста више, што је пожељно јер је основни циљ да се стекне што више знања о карактеристикама, предностима и манама сваког алата. Само након компарације је могуће донети адекватну одлуку о избору алата. У овом прегледу издвојено је 11 алата који су се нашли међу анализираним у најмање два, или више радова. Од ових 11, 7 су искључиво комерцијални, док су осталих 4 првенствено бесплатни – с тим да постоје и плаћене верзије ових алата које пружају већи број функционалности. Алати који нису анализирани у више од једног рада су груписани у посебну категорију и они су у највећем броју бесплатни. Из овога се може закључити да су на тржишту присутнији комерцијални алати, иако укупно гледано постоји више оних бесплатних.

У 5 најчешће анализираних алата налазе се 2 бесплатна и 3 комерцијална. Компаније које стоје иза комерцијалних алата су технолошки гиганти – Мајкрософт (енгл. *Microsoft*), Аj-Би-Ем (енгл. IBM) и Информатика (енгл. *Informatica*). Ова појава има смисла када се узме у обзир да ове компаније углавном нуде софистициране софтверске пакете за управљање и чување (складиштење) података. Мајкрософтов алат, SSIS (*SQL Server Integration Service*), је компонента њиховог *SQL Server Database* софтверског пакета. Аj-Би-Ем-ов *Infosphere datastage* је део *IBM Information Platforms Solutions*, док је једино *Power center* од Информатике самосталан алат – али и он је такође део шире понуде производа за управљање подацима коју ова компанија има. Међутим, алат који је најчешће био анализиран (у 17 радова од 24) је PDI (*Pentaho Data integration*), познат и под именом *Kettle*. Овај алат, који од 2015. године припада фирми *Hitachi Vantara* [26], првенствено је бесплатан. На трећем месту, иза SSIS-а, налази се *Talend Open Studio* који је такође алат отвореног кôда (*open source*) и који је био предмет анализе у 13 од 24 рада. Комплетан преглед ETL алата и број радова у којима су они били анализирани и међусобно поређени дати су у табелама 1 и 2.

У анализираној примарној литератури аутори су у својим истраживањима представљали различите критеријуме за оцењивање и компарацију ETL алата. Они могу бити груписани у пет категорија: перформансе, цена коштања, употребљивост, архитектура и функционалност.

Детаљан преглед критеријума, разврстаних по уоченим категоријама, као и табеле у којима су приказани резултати упоредне анализе ETL алата на основу издвојених радова могу се наћи у [27].

Табела 1. *Преглед поређених алата у примарној литератури*

ETL алат	*Pentaho Data Integration*	*SQL Server Integration Service*	*Talend Open Studio*	*Informatica Power center*	*IBM Infosphere Datastage*	*Clover ETL*
Број радова	17	14	13	9	8	7

Табела 2. *Преглед поређених алата у примарној литератури* (*наставак*)

ETL алат	*SAS Data integration Studio*	*Oracle Warehouse Builder*	*Jaspersoft*	*Oracle Data Dntegrator*	*Ab Initio*	Остали алати
Број радова	4	4	4	3	2	11

3. ОГРАНИЧЕЊА ПРЕГЛЕДА

Методологија извођења овог систематског прегледа литературе одступа од оне која је предложена у [28]. Скоро половина радова је пронађена мануелним претраживањем, што указује на могућност да је изостављен потенцијално значајан број радова који се баве темом компарације и избора оптималног ETL алата. Такође, за неке ручно пронађене радове није могуће гарантовати да су прошли ригорозни процес провере од стране научне заједнице, с обзиром на то да није очигледно да припадају иједном научном часопису (пре свега то су радови [17,24]), иако по другим својим карактеристикама одају утисак научног рада. Даље, постоји ризик од несвесне пристрасности у току селекције литературе јер је првенствено један истраживач спровео избор радова на основу наведених критеријума инклузије и ексклузије. Иста је ситуација и са екстракцијом података, мада број таквих обрађених радова није претерано велик, те ризик од грешке није значајан.

4. ЗАКЉУЧАК

Овај преглед литературе спроведен је у циљу стицања увида у тржишне трендове и утврђивања позиције комерцијалних и бесплатних ETL алата на тржишту.

На основу анализе примарне литературе, издвојено је 11 алата, од којих су 7 комерцијални, а 4 примарно бесплатни (уз постојање плаћених верзија ових алата). Ово указује на већу присутност комерцијалних алата на тржишту. Анализирани радови су предлагали различите критеријуме за компарацију и оцењивање алата, који су у овом раду сврстани у пет категорија: перформансе, цена коштања, употребљивост, архитектура и функционалност. Пожељно би било

више радити на поређењу перформанси алата на практичним примерима и у реалним околностима, као и испитати мишљења и ставове експерата из индустрије, спровођењем анкета на различитим тржиштима.

5. ЛИТЕРАТУРА

[1] H. S. Rose, P. Raibagkar, „A Comparative Study of ETL Tools“, *IJSRD-International Journal for Scientific Research & Development*, том 4, изд. 5, стр. 315–319, 2016.

[2] R. Kimball, J. Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, First. Wiley Publishing, 2004.

[3] A. Kabiri, D. Chiadmi, „Survey on ETL processes“, *Journal of Theoretical and Applied Information Technology XX st Month*, стр. 219–229, 2013.

[4] N. Biswas, A. Sarkar, K. C. Mondal, „Empirical Analysis of Programmable ETL Tools“, у *Communications in Computer and Information Science*, 2019, том 1031, стр. 267–277.

[5] T. A. Majchrzak, T. Jansen, H. Kuchen, „Efficiency evaluation of open source ETL tools“, у *Proceedings of the 2011 ACM Symposium on Applied Computing - SAC '11*, 2011, стр. 287.

[6] J. P. A. Runtuwene, I. R. H. T. Tangkawarow, C. T. M. Manoppo, R. J. Salaki, „A Comparative Analysis of Extract, Transformation and Loading (ETL) Process“, у *IOP Conference Series: Materials Science and Engineering*, Феб. 2018, том 306, изд. 1.

[7] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, S. ben Yahia, „Data quality in ETL process: A preliminary study“, *Procedia Comput Sci*, том 159, стр. 676–687, 2019.

[8] D. Narandžić, S. Ristić, D. Stefanović, T. Lolić, „The Challenge of an Extraction-Transformation-Loading Tool Selection“, у *XIV International Conference on Systems, Automatic Control and Measurements SAUM*, Niš, 2018, стр. 42–45.

[9] V. M. Parra, A. Syed, A. Mohammad, M. N. Halgamuge, „Pentaho and Jaspersoft: A Comparative Study of Business Intelligence Open Source Tools Processing Big Data to Evaluate Performances“, 2016.

[10] M. L. Grecol, „Microsoft SSIS and Pentaho Kettle: A Comparative Study for Three-Tier Data Warehouses“, Statesboro, 2012.

[11] Akshay Dinesh Badgujar, Saurabh Shrikant Kadam, Manasi Mohan Zambare, Shubham Raghavendra Kulkarni, „A Comparative Study: Business Intelligence Tools“, *International Journal of Research in Engineering, Science and Management*, том 5, изд. 1, стр. 98–100, 2022.

[12] V. A. Kherdekar, P. S. Metkewar, „A technical comprehensive survey of ETL tools“, *International Journal of Applied Engineering Research*, том 11, изд. 4, стр. 2557–2559, Март 2016.

[13] M. Patel, D. B. Patel, „Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future“, у *Advances in Intelligent Systems and Computing*, 2021, том 1187, стр. 389–398.

[14] R. Katragadda, S. S. Tirumala, D. Nandigam, „ETL tools for Data Warehousing: An empirical study of Open Source Talend Studio versus Microsoft SSIS“, 2015.

[15] J. Sreemathy, R. Brindha, M. Selva Nagalakshmi, N. Suvekha, N. Karthick Ragul, M. Praveennandha, „Overview of ETL Tools and Talend-Data Integration“, у *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Март 2021, стр. 1650–1654.

[16] G. S, K. Rajesh Dhanani, P. Pankaj Doshi, „DATA ANALYSIS AND ETL TOOLS IN BUSINESS INTELLIGENCE“, *International Research Journal of Computer Science*, том 07, изд. 05, стр. 127–132, Maj 2020.

[17] N. Rodriguez, K. Lawson, E. Molina, J. Gutierrez, „Data Warehousing Tool Evaluation-ETL Focused“, Edinburg, 2011.

[18] I. I. Kholod, M. S. Efimova, S. Ya. Kulikov, „Using ETL Tools for Developing a Virtual Data Warehouse“, 2016.

[19] P. S. Diouf, A. Boly, S. Ndiaye, „Performance of the ETL processes in terms of volume and velocity in the cloud: State of the art“, у *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Нов. 2017, стр. 1–5.

[20] Md. Badiuzzaman Biplob, G. A. Sheraji, S. I. Khan, „Comparison of Different Extraction Transformation and Loading Tools for Data Warehousing“, у *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, Окт. 2018, стр. 262–267.

[21] M. N. Mali, M. Sachinbojewar, „A Survey of ETL Tools“, *International Journal of Computer Techniques*, том 2, изд. 5, стр. 20–26, 2015.

[22] R. Mukherjee, P. Kar, „A Comparative Review of Data Warehousing ETL Tools with New Trends and Industry Insight“, у *2017 IEEE 7th International Advance Computing Conference (IACC)*, Јан. 2017, стр. 943–948.

[23] S. Misra, S. K. Saha, C. Mazumdar, „Performance Comparison of Hadoop Based Tools with Commercial ETL Tools – A Case Study“, у *LNCS*, том 8302, 2013, стр. 176–184.

[24] N. Schmidt, M. Rosa, R. Garcia, E. Molina, R. Reyna, J. Gonzalez, „ETL Tool Evaluation-A Criteria Framework“, Edinburg, 2011.

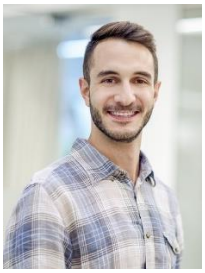
[25] J. Singh, A. Singh, „A comparative Review of Extraction, Transformation and Loading Tools“, *Database Systems Journal*, том 4, изд. 2, стр. 42–51, 2013.

[26] „Hitachi Data Systems Completes Pentaho Acquisition“, Јуни 04, 2015. https://www.hitachivantara.com/en-us/news/in-the-press/2015/gl150604.html (приступљено у августу 2022).

[27] Н. Ковачевић, „Упоредна анализа ETL алата – систематски преглед литературе и студија случаја пројектовања и имплементације складишта података“, Факултет техничких наука, Нови Сад, 2022.

[28] B. A. Kitchenham, S. Charters, „Guidelines for performing Systematic Literature Reviews in Software Engineering“, Авг. 2007.

Кратка биографија:



Никша Ковачевић рођен је у Новом Саду 1997. год. Дипломски рад на Факултету техничких наука из области индустријског инжењерства и инжењерског менаџмента – Предвиђање исхода тениског меча коришћењем техника машинског учења одбранио је 2019. год.
Контакт: niksakovacevic@gmail.com