# Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future


**Monika Patel and Dhiren B. Patel**

**Abstract** ETL is the bedrock of a data warehouse. As the data accelerates in the diversified fields, it is substantive to integrate the data to extract the prerequisites for business advancement and policy formation in the arena of health care, education, smart cities, transportation, and many other areas. ETL (Extract–Transform–Load) is the process used to incorporate a range of data sources in the data warehouse for business intelligence. As per the requirement of data integration and analysis, several categories of ETL tools have developed like code-based, GUI-based, cloud-based, Metadata support, Real-time support, and batch processing. Selecting appropriate ETL tools is a crucial task in any business, with each tool having its dominant and stumbling blocks. This paper aims to first, the study of existing ETL tools and its features, and secondly, how to identify vital functions of the ETL tool for an organization.

**Keywords** ETL · Data warehouse · Business intelligence

## 1 Introduction

ETL is the process used worldwide for developing Data Warehouses. (DW) It is persuasive to think that data warehouse is nothing but Extracting data from various data sources and loading it into it. In the market, a variety of ETL tools is available like commercial, open source, Code-based, GUI-based, cloud-based, and many others. Selecting the appropriate tool is depending on the requirement, type of data, and infrastructure available at the organization. As the data changes, ETL tools are also needed to change as per business and decisions it requires.

M. Patel (✉) · D. B. Patel
Department of Computer Science, Gujarat Vidhyapith, Ahmedabad, Gujarat, India
e-mail: monapatel18r@gmail.com

D. B. Patel
e-mail: dhiren_b_patel@gujaratvidhyapith.org



V. S. Rathore et al. (eds.), *Rising Threats in Expert Applications and Solutions*,
Advances in Intelligent Systems and Computing 1187,
https://doi.org/10.1007/978-981-15-6014-9_45

### *1.1 ETL*

ETL involves three tasks to be performed named Extract, Transform, and Load. Each job has a procedure. Extracting is the process of consolidating data from various sources such as ERP, SAP, and other operational systems into the data warehouse or staging area, ready for the transformation process. The transforming process of ETL involves many subtasks, and those are generally applying business rules, cleaning, filtering, splitting, joining, transporting, and validating the extracted data, to make it ready for loading. Loading process loads data into the data warehouse or other repositories for storage [1].

### *1.2 Data Warehouse*

Data warehouses use for reporting and business analysis. Data warehouses are the central storage for one or more different sources like historical data or current data. Such integrated data used to create reports for analytics for various aspects of organizations. ETL is the process for creating staging and data integration for data warehouses. The staging layer stores raw data collected from each of the distinct sources into the staging area or staging database. The integration layer combined the data from different sources and moved to the database, usually called a data warehouse. In the data warehouse, data is arranging in groups called dimensions, facts, and aggregated facts [2].

## 2 Related Work

Nowadays, a large amount of data created from a large number of devices and other transactions. Such data is usually unstructured or semi-structured. Many tools are available, both open source and commercial, to handle structured data. Traditionally, unstructured data is converted into structured data and then used for analytics. In this time, turning such a massive amount of data into formal is a time-consuming process [3]. To overcome such a problem, authors have proposed a skill-full mechanism to handle unstructured data using a NoSQL database such as MongoDB [4].

In this research, authors have compared various dimensions of some exceedingly used ETL tools such as Informatica, Datastage, AbInitio, Oracle Data Integration, and SSIS. Then the author has also discussed its advantages and disadvantages. The comparison is made based on real-time analysis, language binding, performance optimization, cost, data lineage [2].

This paper focuses on NoSQL systems, which are being simplified and used to implement data warehouse. Previously, the pre-awareness of target system schema is a prerequisite for ETL tools. In contrast, schema-free NoSQL databases do not

have any such requirements. In this work, researchers proposed ETL-based stage for moving a multidimensional theoretical model to a document-based model [5].

In this research paper, the author has reviewed 97 papers to evaluate modern approaches used to employ ETL solutions [6, 7]. Paper highlights the emerging technologies which can be used to develop next-generation ETL tools such as Machine Learning, Artificial Intelligence, and Robotics [8].

In this competitive market, many organizations develop a system without doing the investigation and needs of a company, or without finding benefits and outcomes, companies try to get through development [9]. This paper analyzed two widely used BI tools: Pentaho and Jaspersoft. This research focuses on the ETL, Reporting, and performance while processing with six different sized databases for both tools [10].

This research has been done to compare two ETL tools named SSIS and PDI. PDI is the open-source ETL tool can be used for small organizations with the limited debugging facility. Finally, the choice of selecting the ETL tool depends on the type of project it leads [11].

Data without integration and without stored in a standardized format is the wastage [12]. The ETL process is the main component for making data warehouse. Wrong choices can mislead the organization, so choosing the right ETL tool for DW is crucial. In this research, the analysis of some popular ETL tools with their features is studied [13].

## 3 ETL Tools Categories

Extract, Transform, and Load (ETL) tools facilitate enterprises to make their data available, consequential, and serviceable across unrelated data systems. When selecting the right ETL tool, plenty of options available. In this section of paper, we have identified the categories and their standard features, which we will be useful for selecting the best tool you need. The majority of tools fall in more than one class due to its powerful features.

### *3.1 Batch Processing*

Data gathering before processing in batches is the traditional way of processing. The whole file is received, parsed, validated, cleaned, calculated, aggregated, and then deliver to some system for further evaluation. Modern enterprises can't wait for the collection of data for processing; it requires instant processing of received data. Some other batch processing-based ETL tools are IBM InfoSphere DataStage, SSIS, Informatica, and Oracle Data Integrator [2].

### 3.2 Code-Based/Engine-Based

Code-Based ETL tools are programs, which are compiled and are not writing in a proprietary language. Usually, code generates ETL programs in universal languages like COBOL or C. Engine based are generally proprietary solutions, with their unique data engines designed for performance. They also allow improving the functionality of the product. Both Code-based and Engine-based ETL tools provide excellent performance, but if you don't have the expertise to customize the code, you can choose engine-based ETL tools. Some code-based ETL tools are Oracle Warehouse builder and SSIS [13].

### 3.3 Cloud-Based

Cloud-based ETL tools propose scalability, real-time, streaming data processing, and integrations with a continuously mounting number of data sources. Matilian, Blendo, Stitch, Fivetran, and Alooma are some more examples of cloud-based ETL tools [14].

### 3.4 Open Source

Some people prefer to avail of open-source solutions as it is less expensive than commercial solutions [15]. System integrators, departmental enterprise developers, and mid-market companies most like to use and support open-source ETL tools due to their fewer complex data and reporting requirement. Apache AirFlow, Apache Kafka, Apache NiFi, and Talend Open Studio are some examples of open-source ETL tools [10].

### 3.5 GUI-Based

GUI-based ETL tools are the widely used tools in the market due to the easy user interface of the application. GUI-based tools provide drag and drop facility for data loading and analyzing. Pentaho, Informatica, DataStage, Abinitio are some GUI-based ETL tools [2, 15].

### *3.6 Real-Time*

With the evolution of data and its sources, requires real-time data processing. The demand for real-time data has changed the architecture of ETL. Alooma, Confluent, StreamSets, and Striim are some real-time ETL tools [16].

### *3.7 NoSQL-Based*

ETL process of collecting digital data and converting the information in understandable form like reports is as old as the digital era. With the development of schema-less databases, it becomes complicated to perform the analysis. MongoDB is the most famous NoSQL-based database used widely [17]. Some free/open source and paid versions of MongoDB ETL tools are available. Some other tools which can allow data integration with NoSQL database like MongoDB are MongoSyphon, Transporter, Krawler, Panoply, Stitch, Talend Open Studio, and Pentaho [18, 19].

## 4 Feature Comparison of ETL Tools

We have identified important features for an effective ETL tool based on the survey done by many researches and web sources [2, 11, 13, 20, 21]. Table 1 depicts the comparison of features, major industries are looking for Data Transformation (T), Real-time Integration (R), Real-time Analytics (R), GUI Support (G), Cloud Support(C), Horizontal Scalability (S), Non-RDBMS connections (N), Parallel Processing (P), Metadata Management (M), and Automation (A) features.

In this paper, 15 ETL tools are studied extensively to find out features provided by each tool, and 9 essential elements highlight. In this paper, 9 features are selected and compared against 15 varieties of ETL tools. Table 1 shows that whether the feature

**Table 1** Feature comparison of ETL tools

| Features tools | T | R | G | C | S | N | P | M | A |
|---|---|---|---|---|---|---|---|---|---|
| OWB | Y | N | Y | N | N | N | N | Y | L |
| SSIS | Y | Y | Y | Y | N | N | Y | Y | L |
| Informatica | Y | Y | Y | Y | L | N | Y | Y | L |
| IBM data stage | Y | Y | Y | Y | N | N | Y | Y | L |
| Fivetran | Y | Y | N | Y | N | Y | Y | N | Y |
| Stitch | Y | N | N | Y | N | Y | Y | N | Y |

(continued)

**Table 1** (continued)

| Features tools | T | R | G | C | S | N | P | M | A |
|---|---|---|---|---|---|---|---|---|---|
| AbInitio | Y | Y | Y | N | N | N | Y | Y | L |
| Alooma | Y | Y | N | Y | N | N | N | N | L |
| XPlenty | Y | Y | Y | Y | N | Y | N | N | L |
| NiFi | Y | Y | Y | N | N | Y | N | N | L |
| Talend | Y | Y | Y | Y | N | N | Y | N | N |
| Pentaho | Y | Y | Y | Y | N | Y | N | Y | Y |
| Striim | Y | Y | N | Y | N | Y | N | N | L |
| MongoSyphon | Y | N | N | N | Y | Y | N | N | N |
| Krawler | Y | Y | N | N | Y | Y | Y | N | L |

is present in the tool or not, some features are not fully present or absent in the tool. L represents limited coverage of features in the specified tool [22–24].

Oracle Warehouse Builder provides end to end data integration with data transformation using PL/SQL functions, procedures, and packages and handles erroneous data. Microsoft SSIS is a data warehousing tool that provides data transformation using fuzzy lookup and grouping functionality with message-based real-time data integration. Informatica is a tool used for connecting and fetching data from various sources and it provides advanced data transformation using parsing. IBM InfoSphere DataStage is the ETL platform that offers data integration across many enterprises. DataStage offers real-time data integration and automates monitoring. Fivetran is a tool that helps analysts by using orchestrate SQL-based data transformation and real-time monitoring and data integration and also provides non-RDBMS connectors for NoSQL databases, such as MongoDB. Stitch is a powerful ETL tool, which integrates the organization's data into the central data warehouse [25–27]. AbInitio is a powerful, GUI-based ETL tool that performs parallel processing for data management. It provides limited automation among its features.

Alooma provides real-time data streaming with cloud support and uses code engines for data transformation [28]. Xplenty is the no coding GUI environment which performs data integration and processing from various data sources. It provides real-time data chunker and masking and also automates server configuration ratio. Apache NiFi supports directed graphs for data transformation. Talend Open Studio is an ETL tool that connects a large number of databases and applications and performs extensive transformation using the Native Code. Pentaho Data Integration creates self-documented data pipelines for processing, and it uses a spoon tool for data transformation. Striim is the real-time data analytics platform and reduces architectural complexity. MongoSyphon is an open-source ETL tool that transforms data into documents that store in document-oriented databases. Krawler is the minimalist geospatial ETL tool. It automates the extraction process [29–31].

**Table 2** Choosing ETL tool for an organization

| Tool | Type of organization | Usage | Organization size |
|---|---|---|---|
| Fivetran | • Internet<br>• Education | 80% | Mid-Market (51–1000 Emp) |
| Xplenty | • E-Learning<br>• Marketing<br>• Advertising | 54% | Small business (<50 Emp)<br>Mid-Market (51–1000 Emp) |
| Stitch | • Computer software<br>• Information technology services | 55.3% | Small business (<50 Emp) |
| Pentaho | • Telecommunications<br>• Financial services | 33.3% | Suitable for all |
| AbInitio | • Education<br>• Service oriented<br>• Manufacturing | – | Small business (>50 Emp) |
| Informatica | • Healthcare<br>• Retail<br>• Finance | 84.6% | Enterprise (>1000 Emp) |
| Alooma | • Education<br>• Consumer electronics | 25% | Mid-Market (51–1000 Emp) |
| Apache NiFi | • Health and wellness<br>• Telecommunication<br>• Marketing | 18.2% | Enterprise (>1000 Emp) |
| Striim | • Environmental services | 82% | Enterprise (>1000 Emp) |
| Talend Open Studio | • Information technology and services<br>• Media | 33.3% | Enterprise (>1000 Emp) |
| IBM Infosphere DataStage | • Banking | 87.5% | Enterprise (>1000 Emp) |

## 5 Review Findings

In this section, review findings are provided based on a study performed on distinct 15 ETL tools and their feature comparison [32]. This paper covers features that are essential for any ETL tool [33]. As per the study, we have found out that which ETL tool provides maximum features from the given list. Figure 1 shows ETL tools used by small businesses, Mid-market, and Enterprises. While Selecting an ETL tool, the organization must identify its essential features for fulfillment. Based on the feature present in the ETL tool, an organization can select the tool for data integration. This paper also highlights organization types and sizes suitable for various ETL tools. This study can be handy for an organization to choose the appropriate ETL tool.

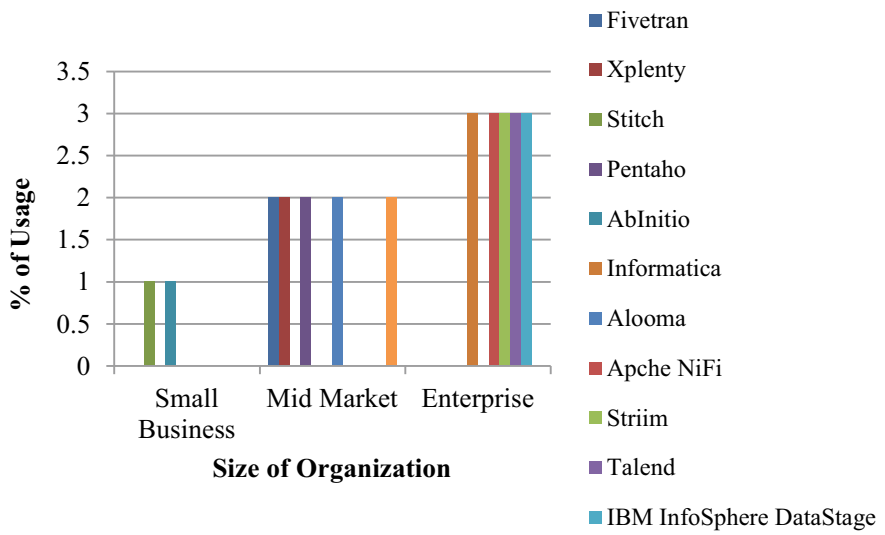



**Fig. 1** ETL tools for the size of the organization

## 6 Conclusion

This paper focuses on categories of ETL tools and products available for each type. Many tools can fall into more than one class due to its features and implementation techniques. Choosing the appropriate ETL tool for your organization is crucial in any organization. Selected ETL tool should not be highly configured then actually required. If you don't need real-time updates and does not need to handle stream datasets, you can choose any simple tool that can fit for your requirement. If you deal with a large amount of data or streaming data, you can build your ETL tool based on open-source technology. Based on a study finding essential criteria for the selection process of ETL tools are many. It categorizes as a type of organization, type of use, existing data storage, connectivity, real-time processing, data integration options, visualization support, and cost. There is a variety of ETL tools available in the market. However, there is still scope of improvement, especially in open-source technology for small scale businesses to provide more features in one tool with affordable cost.

## 7 Future Enhancement

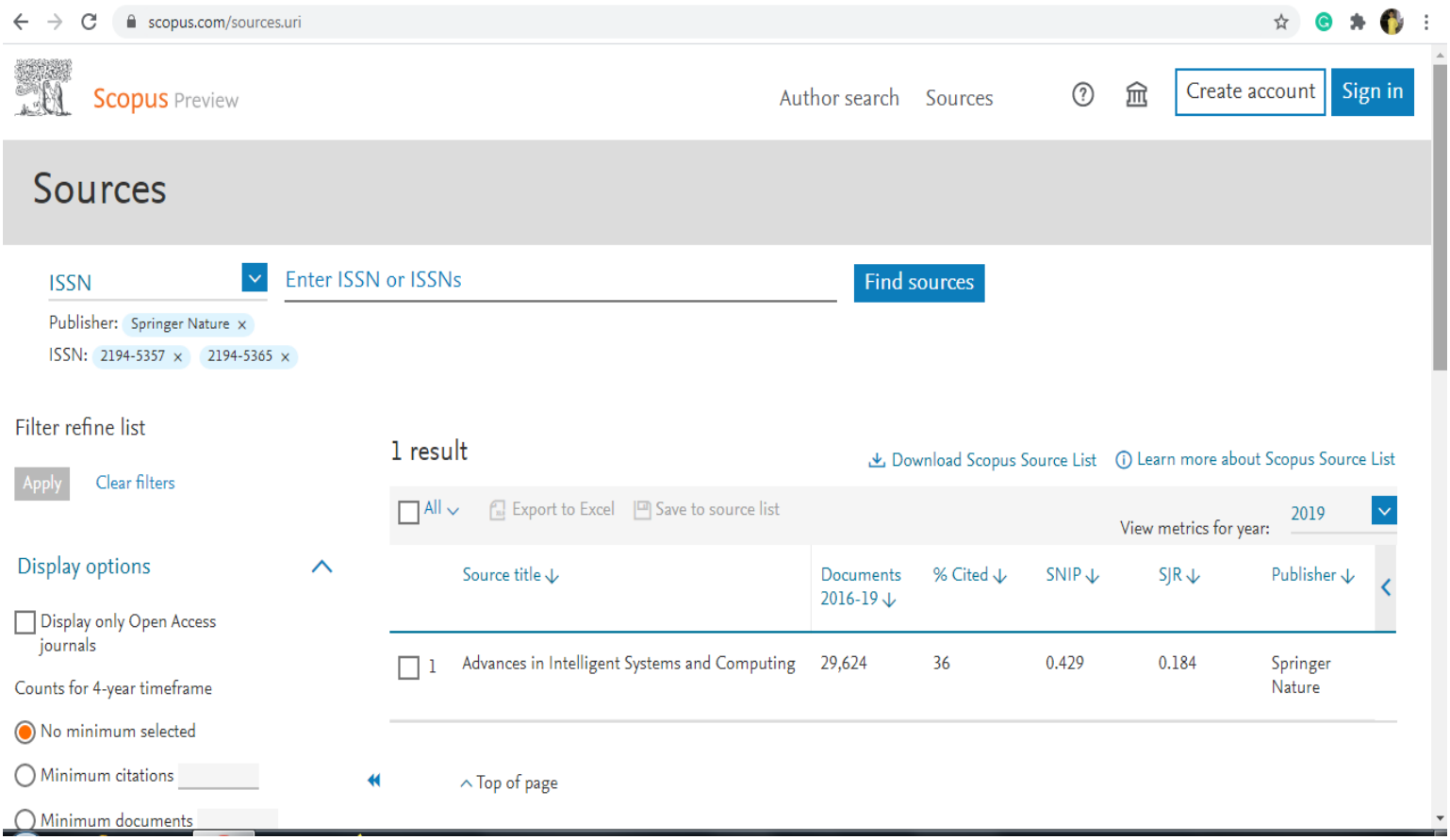
This paper covered various categories of ETL tools for diverse requirements. With that, this paper has also covered that no ETL tool provides all fundamental features. The demand for a new ETL tool is elevating, which can perform better in terms

of cost, transformation, non-RDBMS connections, parallel processing for such a massive amount of data. In the future, one can address one of the following challenges: automation, non-DBMS connectors, metadata management, and parallel processing.

## References


1. C. Thomsen, ETL. Springer International Publishing AG, Part of Springer Nature (2018)
2. R. Mukherjee, P. Kar, A comparative review of data warehousing ETL tools with new trends and industry insight, in *IEEE 7th International Advance Computing Conference* (2017), pp. 943–948
3. J. Awiti, E. Zimányi, An XML interchange format for ETL models, in *New Trends in Databases and Information Systems. ADBIS 2019,* ed. by T. Welzer et al. Communications in Computer and Information Science, vol. 1064 (Springer, 2019)
4. R. Aluvalu, M. Jabbar, Handling data analytics on unstructured data using MongoDB, in *Smart Cities Symposium* (2018), pp. 1–5
5. R. Yangui, A. Nabli, F. Gargouri, *ETL Based Framework for NoSQL Warehousing*, Lecture Notes in Business Information Processing (Springer, Cham, 2017)
6. J. Wang, W. Zhao, T. Fan, S. Yang, H. Lv, An improved join free snowflake schema for ETL and OLAP of the data warehouse. Concurr. Comput. Pract. Exper. (2019)
7. N. Biswas, A. Sarkar, K.C. Mondal, Empirical analysis of programmable ETL tools, in *Computational*, ed. by *Intelligence, Communications, and Business Analytics, CICBA 2018,* ed. by J. Mandal, S. Mukhopadhyay, P. Dutta, K. Dasgupta. Communications in Computer and Information Science, vol. 1031 (Springer, Singapore, 2019)
8. J. Nwokeji, F. Aqlan, A. Apoorva, A. Olagunju, Big Data ETL implementation approaches: a systematic literature review, in *Conference of Software Engineering and Knowledge Engineering* (2018)
9. S.M.F. Ali, R. Wrembel, Towards a cost model to optimize user-defined functions in an ETL workflow based on user-defined performance metrics, in *Advances in Databases and Information Systems. ADBIS 2019,* ed. by T. Welzer, J. Eder, V. Podgorelec, A. Kamišalić Latifić. Lecture Notes in Computer Science, vol. 11695 (Springer, Cham, 2019)
10. V. Para, A. Mohammad, A. Syed, M. Halgamuge, Pentaho and Jaspersoft: A comparative study of business intelligence open source tools processing big data to evaluate performances. Int. J. Adv. Comput. Sci. Appl. (2016)
11. A. Amine, R. Daoud, B. Bouikhalene, Efficiency comparison and evaluation between two ETL extraction tools. Indones. J. Electric. Eng. Comput. Sci. 174–181 (2016)
12. J. Awiti, A. Vaisman, E. Zimányi, From conceptual to logical ETL design using BPMN and relational algebra, in *Big Data Analytics and Knowledge Discovery, DaWaK 2019,* ed. by C. Ordonez, I.Y. Song, G. Anderst-Kotsis, A. Tjoa, I. Khalil. Lecture Notes in Computer Science, vol. 11708 (Springer, 2019)
13. S. Yousuf, S. Rizvi, A comparative study of ETL tools, https://www.academia.edu/354387
14. P. Diouf, A. Boly, S. Ndiaye, Performance of the ETL processes in terms of volume and velocity in the cloud: state of the art, in *4th IEEE International Conference on Engineering Technologies and Applied Sciences* (2017)
15. J. Chakraborty, A. Padki, S. Bansal, Semantic ETL—state-of-art and open research challenges, in *IEEE 11th International Conference on Semantic Computing. San Diego, CA* (2017)
16. P. Diouf, A. Boly, S. Ndiaye, Variety of data in the ETL processes in the cloud: state of the art, in *IEEE International Conference on Innovative Research and Development, Bangkok, Thailand* (2018)
17. I. Oditis, Z. Bicevska, J. Bicevskis, G. Karnitis, Implementation of NoSQL-based data warehouses. Baltic J. Modern Comput. **6**, 45–55 (2018)

18. Panoply Blog Home page, https://blog.panoply.io/top-9-mongodb-etl-tools
19. M. Moly, O. Roy, A. Hossain, An advanced ETL technique for error-free data in data warehousing environment. Int. J. Sci. Res. Eng. Trends, 554–558 (2019)
20. A. Pall, J. Singh, ETL Methodologies, limitations, and framework for the selection and development of an ETL tool. Int. J. Res. Eng. Appl. Sci. **6** (2016)
21. M.B. Biplob, G.A. Sheraji, S.I. Khan, Comparison of different extraction transformation and loading tools for data warehousing, in *2018 International Conference on Innovations in Science, Engineering, and Technology* (2018), pp. 262–267
22. R. Katragadda, S.S. Tirumala, D. Nandigam, ETL tools for data warehousing: an empirical study of open source Talend Studio versus Microsoft SSIS
23. I.I. Kholod, M.S. Efimova, S.Y. Kulikov, Using ETL tools for developing a virtual data warehouse, in *2016 XIX IEEE International Conference on Soft Computing and Measurements* (2016), pp. 351–354
24. Gartner Peer Insights, https://www.gartner.com/reviews/market/data-integration-tools. Accessed 15 Sep 2019
25. H. Mallek, F. Ghozzi, O. Teste, F. Gargouri, BigDimETL with NoSQL database, in *22nd International Conference on Knowledge-based and Intelligent Information & Engineering Systems* (2018)
26. B. Pan, G. Zhang, X. Qin, Design and realization of an ETL method in business intelligence project, in *3rd IEEE International Conference on Cloud Computing and Big Data Analytics* (2018)
27. B. Nabila, B. Ladjel, K. Selma, Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service. Cluster Comput. **16**(4), 915–931 (2013)
28. Data Warehouse Guide, https://panoply.io/data-warehouse-guide/etl-tools/. Accessed 2 Sep 2019
29. MongoDB Home page, https://www.mongodb.com/use-cases/real-time-analytics. Accessed 5 Sep 2019
30. S.M.F. Ali, R. Wrembel, From conceptual design to performance optimization of ETL workflows: current state of research and open problems. VLDB J. (2017)
31. B. Nabila, B. Ladjel, G. Laurent, ETL processes in the era of variety. Large-Scale Data- and Knowl.-Centered Syst. **39**, 98–129 (2018)
32. G2, https://www.g2.com/search?utf8=%E2%9C%93&query=ETL+tools. Accessed 17 Sep 2019
33. J. Smith, M. Rege, *The Data Warehousing Evolution: Where's it headed next?* (ACM, ICCDA, 2017)


scopus.com/sources.uri
Scopus Preview
Author search
Sources
Create account
Sign in
Sources
ISSN
Enter ISSN or ISSNs
Find sources
Publisher: Springer Nature
ISSN: 2194-5357 2194-5365
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2019
Source title
Documents 2016-19
% Cited
SNIP
SJR
Publisher
1
Advances in Intelligent Systems and Computing
29,624
36
0.429
0.184
Springer Nature
Top of page