Fuel Cells Bulletin
ISSN: 1464-2859

How Companies Use ETL Tools to Move Data Between Systems

Luis Eduardo Muñoz Guerrero
Universidad Tecnológica de Pereira, Colombia
Email: lemunozg@utp.edu.co
ORCID: https://orcid.org/0000-0002-9414-6187

Abstract

Many companies today have a lot of data. This data is stored in different systems and databases, and sometimes it needs to be moved from one place to another. ETL, which stands for Extract, Transform, and Load, is a process that helps companies do this. The basic idea behind ETL is straightforward: data is first extracted from one or more source systems, then transformed into a format that is suitable for analysis or reporting, and finally loaded into a destination system such as a data warehouse. Even though this process has been around for several decades, it remains one of the most commonly used approaches for data integration in organizations of all sizes.

In this paper, we look at how ETL works in practice and why so many companies continue to rely on it. We describe each of the three main steps of the ETL process in some detail, explaining what happens during extraction, what kinds of transformations are typically applied to data, and what the loading step involves. We also discuss the difference between full loads and incremental loads, which is an important practical consideration for organizations that run ETL jobs on a regular schedule.

Additionally, the paper provides an overview of some of the most widely used ETL tools currently available, covering both commercial options such as Informatica PowerCenter and Microsoft SQL Server Integration Services (SSIS), as well as popular open source alternatives like Talend Open Studio, Apache NiFi, and Pentaho Data Integration. We compare these tools in terms of their type, primary use case, and cost, which we believe is useful information for organizations that are trying to choose a tool.

We also dedicate a section to discussing the main challenges and limitations of ETL. These include performance issues that arise when processing large volumes of data, data quality problems inherited from source systems, the ongoing maintenance burden that comes with keeping ETL pipelines up to date as source systems change, and the fundamental limitation of traditional batch-oriented ETL in scenarios where real-time data is required.

Our findings suggest that ETL remains a relevant and practical solution for a wide range of data integration scenarios, particularly in organizations where data is updated on a daily or weekly basis and where real-time processing is not strictly necessary. While newer approaches such as ELT and streaming data pipelines are gaining adoption, ETL continues to be the default choice in many enterprise environments, especially those with existing investments in data warehouse infrastructure. We conclude that understanding ETL is still important for data engineers and database professionals, even as the data landscape continues to evolve.

Keywords: *ETL, data warehousing, data integration, data transformation, business intelligence, data quality, ETL tools.*

1. Introduction

Nowadays, companies have a lot of data. This data comes from many different sources, like sales systems, customer databases, and online platforms. Because the data comes from so many places, it is often stored in different formats and different systems. This can be a problem when a company wants to use all of its data together for reporting or analysis.

ETL is a process that was created to solve this problem. ETL stands for Extract, Transform, and Load. The idea is simple: first you take the data from where it is stored (extract), then you change it into the format you need (transform), and finally you put it somewhere useful like a data warehouse (load). This process has been used in companies for many years and is still very common today.

The goal of this paper is to explain how ETL works and to talk about why companies use it. We will also look at some popular ETL tools and discuss whether ETL is still a good option for companies today. This paper is organized as follows:

Section 2 gives some background on data and databases, Section 3 explains the ETL process in more detail, Section 4 talks about ETL tools, Section 5 discusses some challenges, and Section 6 is the conclusion.

2. Background

Companies have been collecting data for a very long time. In the early days of computing, data was stored on paper or in simple files on computers. As time went on, database management systems (DBMS) were developed, which made it easier to store and organize large amounts of data. Today, companies use many different types of databases, including relational databases like MySQL and PostgreSQL, and also newer types like NoSQL databases.

According to Inmon (1992), who is often called the father of the data warehouse, companies need a central place to store all their historical data so that they can do analysis and make decisions. This idea is still true today. A data warehouse is basically a big database that is designed specifically for analytical purposes, as opposed to operational databases that are used for day-to-day transactions.

The problem with having data in many different places is that it is hard to get a full picture of what is happening in the company. For example, if the sales data is in one system and the customer data is in another system, it is difficult to combine them without some kind of data integration tool. This is where ETL comes in.

A data warehouse is a type of database that is designed for storing and analyzing large amounts of data. Unlike regular databases that are used for everyday operations, a data warehouse is optimized for reading and querying data. Kimball and Ross (2002) describe the data warehouse as a place where data from different source systems is brought together in a consistent format so that analysts and business users can query it easily. The relationship between ETL and data warehouses is very direct: ETL is the process that fills the data warehouse with data. Without ETL, there would be no good way to get data from the operational systems into the data warehouse in a consistent and reliable way.

3. The ETL Process

The ETL process is made up of three main steps: Extract, Transform, and Load. Each step has a specific purpose and together they allow companies to move data from one system to another in a controlled way. Table 1 below summarizes each step, what it does, and some of the common problems that can happen.

Step	What It Does	Common Problems
Extract	Reads data from source systems like databases, files, or APIs	Source systems can be slow or unavailable
Transform	Cleans and converts data into the required format	Data quality issues, inconsistent formats
Load	Writes the processed data into the destination warehouse	Long load times, integrity violations

Table 1. Summary of the three ETL steps and their common problems.

The first step is extraction. This means taking data from one or more source systems, which can be relational databases, flat files like CSV or Excel files, web services, APIs, or even email systems. One important thing to consider during extraction is how often the data should be extracted. Some companies do this once a day, while others do it more frequently depending on how fresh the data needs to be.

Vassiliadis et al. (2002) explain that extraction can be done in two main ways: full extraction, where all the data is extracted every time, and incremental extraction, where only the data that has changed since the last extraction is taken. Incremental extraction is more efficient but also more complex to implement.

The second step is transformation. This is where the data is cleaned and converted into the format needed for the destination system. Transformation can include many different types of operations such as changing date formats, combining fields from different tables, calculating new values, or removing duplicate records. Transformation is often considered the most

complex part of ETL because data from different source systems is often in different formats and may have different conventions.

Data quality is also addressed during transformation. This includes handling missing values, correcting obviously wrong values, and applying business rules. For example, a business rule might say that a customer's age cannot be negative, so any negative age values should be treated as errors. Improving data quality during transformation is a critical concern, as noted by Timmerman and Bronselaer (2019), who argue that data quality problems must be identified and resolved as early as possible in the data pipeline.

The last step is loading. This means putting the transformed data into the destination system, which is usually a data warehouse. There are two main approaches: a full load, where the destination is completely replaced with new data, and an incremental load, where only new or changed records are added. Most organizations end up using incremental loads for their regular ETL jobs and only do full loads occasionally. According to Kimball and Ross (2002), the loading process should also maintain referential integrity in the destination database to make sure that all relationships between tables remain valid after the load.

4. Common ETL Tools

There are many tools available for doing ETL. Some of them are commercial products that cost money to buy, while others are open source and free to use. Table 2 shows a comparison of some of the most common ETL tools, including their type, main use case, and cost.

Tool	Type	Main Use Case	Cost
Informatica PowerCenter	Commercial	Enterprise data integration	Paid
Microsoft SSIS	Commercial	Windows/SQL Server environments	Paid
Talend Open Studio	Open Source	General ETL workflows	Free
Apache NiFi	Open Source	Data flow automation	Free
Pentaho (Kettle)	Open Source	Small to mid-size projects	Free

Table 2. Comparison of common ETL tools.

Informatica PowerCenter is one of the most widely used commercial ETL tools. It provides a graphical interface for designing ETL workflows and supports connections to many different types of source and destination systems. It is known for being reliable and having good performance, but it is also quite expensive. Microsoft SQL Server Integration Services (SSIS) is another popular option, widely used in organizations that already use Microsoft products.

On the open source side, Apache NiFi provides a web-based interface for designing data flows and supports many different types of connections. Talend Open Studio and Pentaho Data Integration are also popular choices that are free to use and have active communities. Patel and Patel (2020) provide a literature review of the progressive growth of ETL tools over time, observing that the variety of available tools has expanded considerably as the complexity and scale of enterprise data environments have increased.

Khine and Wang (2018) note that with the rise of big data technologies like Hadoop and Spark, there has also been a trend toward using these platforms for ETL processing. This allows companies to process much larger volumes of data than traditional ETL tools can handle. However, these technologies are more complex to set up and require specialized skills. Qaiser et al. (2023) also compared several ETL tools in the context of big data analytics and found that the choice of tool depends heavily on the organization's existing infrastructure and the volume of data being processed.

5. Challenges and Limitations

While ETL is a widely used and generally effective approach for data integration, it does have some challenges and limitations. One challenge is performance. When there is a lot of data to process, ETL jobs can take a very long time to

complete. This is especially problematic if the ETL needs to be done within a specific time window, for example overnight before business users arrive in the morning.

Another challenge is data quality. As mentioned earlier, data from different source systems can have quality problems like missing values, inconsistent formats, or duplicate records. As Rahm and Do (2000) point out, data quality problems are often not discovered until the ETL process is actually running, which can cause unexpected failures. Jouini and Laga (2019) also examined this issue and found that tracking data quality defects throughout the ETL pipeline is essential for ensuring the reliability of the resulting data warehouse.

Maintenance is also a challenge. ETL processes need to be maintained over time as source systems change. If a source system adds a new field or changes the format of an existing field, the ETL code needs to be updated accordingly. In large organizations with many ETL processes, this maintenance work can take up a significant amount of time for the data engineering team. Nwokeji et al. (2018), in their systematic literature review of ETL implementations, found that maintenance and adaptability to changing source systems are among the most frequently cited problems by practitioners.

Finally, ETL has historically been a batch process, meaning that data is only updated periodically rather than in real time. For many use cases this is fine, but for applications that need up-to-date data at all times, traditional ETL may not be suitable. Alkhalil et al. (2019) discuss how the evolution toward real-time ETL architectures addresses this limitation, though they note that real-time integration introduces its own set of technical and operational complexities.

6. Conclusion

In this paper, we have provided an overview of the ETL process and how it is used by companies to move data between systems. We explained the three main steps of ETL: extract, transform, and load. We also discussed some of the most common ETL tools and described some of the challenges associated with ETL.

ETL is an important part of how many companies manage their data. Even though it is a technology that has been around for many years, it is still widely used today and continues to be relevant. The basic idea of extracting data from source systems, transforming it, and loading it into a data warehouse has not changed much over the years, even if the tools and technologies used to do it have evolved.

In conclusion, ETL is a useful and practical approach for data integration that works well for many companies. Future work could look at comparing different ETL tools in more detail or studying how ETL is being used together with newer technologies like cloud computing.

References

1. Alkhalil, A., Khaddaj, S., & Elmorshedy, M. (2019). The evolution of ETL architecture: From traditional data warehousing to real-time data integration. World Journal of Advanced Research and Reviews, 1(3), 073–084. https://doi.org/10.30574/wjarr.2019.1.3.0033
2. Bala, M., Boussaid, O., & Alimazighi, Z. (2017). A fine-grained distribution approach for ETL processes in big data environments. International Journal of Decision Support System Technology, 8(4), 50–69.
3. Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An Overview of Business Intelligence Technology. Communications of the ACM, 54(8), 88–98. https://doi.org/10.1145/1978542.1978562
4. Inmon, W. H. (1992). Building the Data Warehouse. John Wiley & Sons.
5. Inmon, W. H., Strauss, D., & Neushloss, G. (2008). DW 2.0: The Architecture for the Next Generation of Data Warehousing. Morgan Kaufmann.
6. Jouini, K., & Laga, H. (2019). Data quality in ETL process: A preliminary study. Procedia Computer Science, 159, 676–683. https://doi.org/10.1016/j.procs.2019.09.223
7. Khine, P. P., & Wang, Z. S. (2018). Data Lake: A New Ideology in Big Data Era. ITM Web of Conferences, 17, 03025. https://doi.org/10.1051/itmconf/20181703025
8. Kimball, R., & Ross, M. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (2nd ed.). John Wiley & Sons.

9. Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). The Data Warehouse Lifecycle Toolkit. John Wiley & Sons.
10. Mondal, K. C., Biswas, N., & Saha, S. (2020). Role of machine learning in ETL automation. In Advanced Techniques for IoT Applications. Springer Singapore, pp. 172–182.
11. Mhon, G. G. W., & Kham, N. S. M. (2020). ETL pre-processing with multiple data sources for academic data analysis. In IEEE Conference on Computer Applications (ICCA), pp. 1–5.
12. Nwokeji, J. C., Aqlan, F., Apoorva, A., & Olagunju, A. (2018). Big Data ETL Implementation Approaches: A Systematic Literature Review. In Proceedings of the IEEE International Conference on Big Data, pp. 2443–2452.
13. Patel, M., & Patel, D. B. (2020). Progressive growth of ETL tools: A literature review of past to equip future. In Rising Threats in Expert Applications and Solutions, pp. 389–398.
14. Qaiser, A., Farooq, M. U., Mustafa, S. M. N., & Abrar, N. (2023). Comparative analysis of ETL tools in big data analytics. Pakistan Journal of Engineering and Technology, 6(1), 7–12.
15. Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.
16. Saranya, N., Brindha, R., Aishwariya, N., Kokila, R., & Matheswaran, P. (2024). An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing. Journal of Big Data, 6(1). https://doi.org/10.32604/jbd.2024.055252
17. Seenivasan, D. (2023). Exploring popular ETL testing techniques. International Journal of Computer Trends and Technology, 71(2), 32–39. https://doi.org/10.14445/22312803/IJCTT-V71I2P106
18. Simitsis, A., & Vassiliadis, P. (2008). A Methodology for the Conceptual Modeling of ETL Processes. In Advances in Databases and Information Systems. Springer, Berlin.
19. Timmerman, Y., & Bronselaer, A. (2019). Measuring data quality in information systems research. Decision Support Systems, 126, 113138. https://doi.org/10.1016/j.dss.2019.113138
20. Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. In Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP (DOLAP '02), pp. 14–21. ACM. https://doi.org/10.1145/583890.583893