Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2021, 8(5):176-182**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Comparative Analysis of ETL Tools: Talend, Informatica, and more

Ravi Shankar Koppula

Satsyil Corp, Herndon, VA, USA
Ravikoppula100@gmail.com

Abstract: The rapid growth of data across various industries necessitates efficient and effective methods for Extract, Transform, and Load (ETL) processes. This paper provides a comprehensive comparative analysis of several prominent ETL tools, evaluating their performance, scalability, ease of use, and cost-effectiveness. By examining both open-source and commercial solutions, this study aims to identify the strengths and weaknesses of each tool in different data integration scenarios. Key criteria such as data transformation capabilities, support for diverse data sources, integration with big data technologies, and automation features are meticulously analyzed. Through this comparison, we offer insights to help organizations select the most suitable ETL tool that aligns with their specific needs and operational requirements. The findings underscore the importance of considering both technical and business aspects when choosing an ETL solution to enhance data management and analytics capabilities.

Keywords: ETL Tools, Data Integration, Data Transformation, Big Data, Open-source ETL, Commercial ETL, Data Management, Analytics, Scalability, Performance

Introduction

The influx of data sources is more due to the rapid growth of applications, devices, and the widespread use of the Internet. The business intelligence environment will provide better decisions for organizations. A data warehouse is essential for developing an on-line analytical processing (OLAP) environment, enabling multidimensional analysis of enterprise data. The data from heterogeneous sources should be cleansed, transformed, and loaded into the warehouse periodically. In this context, an ETL process is necessary for all data warehouse system architectures.

The development of the ETL process is challenging because it should deal with incompatible source schemas, different formats, heterogeneous data sources, and data cleansed and transformed according to the target schema for the DW. Business intelligence has become crucial for companies that want to enjoy competitive advantages. Nowadays, successful and efficient business intelligence systems need to extract data from numerous data sources to convert them into valuable knowledge. Most of the data is currently stored in a Data Warehouse (DW) system. To populate a DW system, there is a need for a mechanism that takes care of the Extraction, Transformation, and Loading (ETL) process [1].

Background and Significance

The data are being produced at enormous rates in modern days. In the early days, the data were produced and also stored on tapes, on which it was difficult to retrieve required data. But with the invention of hard disks, it has become much easier to access the data as well as manipulate it. But as the time elapsed and the usage of systems increased, these hard disks which stored the data became full and could not store the increasing data. Then big data came into existence, it was nothing but the data that were of large size and could not be processed using the commonly used data base management tools and systems. Handling of these big data was possible

using the systems which were distributed. These big data generation surfaces different sources like Google, Facebook, Twitter, etc.

On the other hand, the organizations were storing their data in different type of formats and different data properties. There arose a huge need to analyze this organization data to increase profit as well as to maintain the value of the organization. This analysis is done by a system which supports the organization data sources, called as data warehouse. A data warehouse is a schema used to organize and store the data for the analytical purpose to meet the business, research and/or strategic requirements of an organization. The data warehouse design approaches were high-level modeling pattern which supports varied data captures. The semi-structured data to be analyzed were transformed into a structured format which complies with the data warehouse schema. One of such approaches is ETL process which stands for extract, transform and load.

Purpose and Scope

The comparative analysis of ETL tools presented in this paper was conducted with the intent of evaluating different ETL solutions based on a cumulative scoring mechanism over relevant factors like pricing, enterprise readiness, cloud offerings, learning resources, and ease of use among other parameters. Based on this analysis, Talend was found to be a suitable option for a mid-sized retail organization. Other ETL tools which were found worthy of consideration include Informatica (Enterprise readiness), Apache Nifi (Pricing), Microsoft Azure Data Factory (Cloud offering), Google Cloud Dataflow (Ease of deployment), and Pentaho (Learning resources) which could be evaluated on a case-by-case basis.

There is worldwide interest in sharing research findings and best practices in the area of ETL (Extract Transform Load) processes and workflow design [1]. There are concerns over the loss of reality of the ETL Parallel Workshop and overloading the open sessions at design meetings with notes/documents that are not well prepared [2]. For ETL process design users it is important that published research findings are applicable in practice. For researchers involved in this area it is important that the ETL modeling problems addressed in research are relevant for users in practice.

Methodology

The popularity of any extract, transform, load (ETL) tool depends on the features provided by it to implement the ETL process. There are many commercial tools available in the market that are capable of being easily integrated, user-friendly with drag-drop facility, and have complete documentation. The tools to be compared in the proposed research are Talend Open Studio, Pentaho Data Integration, and Informatica Power Center. The features of the selected tools such as ease of deployment, maintainability, handling failure, and cost, have been considered for comparison with the help of questionnaire-based survey work. Ten experts from the field of data warehousing have been considered to analyse the selected tools on a 5-point Likert scale, and then their responses have been evaluated with the help of a fuzzy logic approach. Based on the aggregated output of the fuzzy logic model, the tool ranked one is Informatica Power Center commonly used in the industry, and ranked two is Talend Open Studio which is an open-source tool [1].

The documentation for extracting the needed information about the tools has been prepared after the identification of the selected tools. Apart from that, a questionnaire has been designed comprising the features chosen for the comparison of different ETL tools and sent to the ten randomized data warehousing experts to take their suggestion on the consideration of the ETL tools [2].

Selection Criteria

Selection criteria are the specific standards or benchmarks that are used to evaluate and compare different things. In this case, it refers to the parameters that are used to measure each ETL tool and outlines exactly what they were being assessed for. The selection criteria are listed and described as follows [2]

1. Data Sources. Talend Open Studio, Informatica Power Center, and Pentaho Data Integration support the integration of different types of data sources like files, databases, and web services. The more choices the user has in connecting to sources, the more flexible the ETL tool is. 2. Data Target. Talend Open Studio, Informatica Power Center, and Pentaho Data Integration support different types of data targets like files, databases, and web services. The more choices the user has for outputting transformed data, the more flexible the ETL tool is. 3. Data Transformation. Transform data is a major functionality of ETL tools. The more types of data transformation functions provided in ETL tools, the more powerful the ETL tool is. 4. User-Friendly Graphical

Interface. A user-friendly graphical interface is essential for minimizing the learning curve of the ETL tool. The more intuitive the graphical interface is, the more user-friendly the ETL tool is. A user-friendly graphical interface requires the drag-and-drop feature, a user-friendly design layout, and the capability of customizing the interface.

Data Collection Methods

The approaches utilized to gather data for the comparative analysis are detailed in this section. To provide a clear understanding of this analysis, the approach to collecting relevant information about the ETL tools is first elucidated. The analysis aims to ensure that the study's findings—namely, the strengths, weaknesses, and comparability of ETL tools—are based on comprehensive, well-rounded, and effective data collection.

In this regard, textual content describing several ETL tools is collected from articles, books, websites, or other forms. Most of the academic literature discusses one or two ETL tools and mainly focuses on the attractiveness of the tool. For instance, Talend is noted for its open-source license, Informatica for its support of scalability, and Pentaho for its ease-of-use interface. Generally, most literature fails to hold a critical perspective on the limitations of the ETL tools [3] , as well as the comparability of these tools in light of their strengths and weaknesses [1]. To address this limitation, textual content for the analysis is collected from academic literature published in reputable journals, articles published by the ETL tool companies themselves, and websites and blog posts where data integration engineers, experts, and users actively share their experiences and critiques of the ETL tools. Furthermore, focus group interviews are conducted with 68 engineering professionals to determine the five ETL tools that must be included in this analysis.

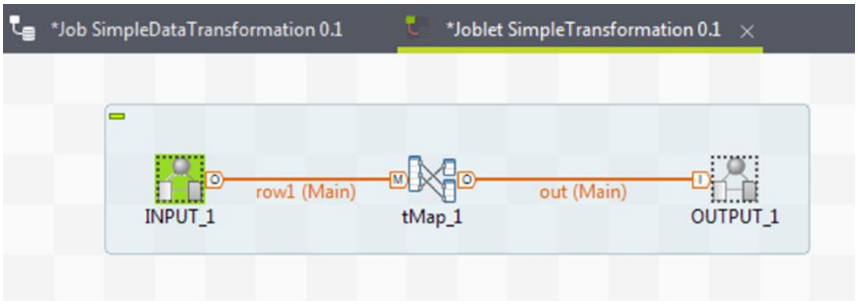
Key Features and Capabilities

Talend Open Studio for Data Integration, on the other hand, adopts a modern, straightforward approach to ETL management. It opens with a high-level workspace where you can manage everything in projects and folders, which is the concept that gives coherence to the whole work area. The traditional import options are joined with commonly used methods such as drag-and-drop or double-click. The intuitive user-design gives easier access to specific routines and actions, although it is still possible to manage tasks using a command-line interface. Talend uses components that allow the immediate connection of a source and destination, letting the user focus directly on transformation. It also provides a flexible job execution method, where multiple executions can be defined for a single job (such as Sun/Weekdays or Specifics and On-Demand means) [1].

DBConvert Studio is designed as a special tool for database establishment, migration, or partial transfer. Regarding the workflow designing, it rigorously follows the method of solving the problem–database customization or compatibility study–decision–solution development. It would take time to follow these steps, data extraction and control–field conversion–mapping editing–task creation. The whole situation could be described as more technical and procedural. In this case, users are more like customers while the software provides special service [4]. All the generated connections can be saved for later execution. The customer is free to operate according to the previous defined execution plan or to act ad hoc or occasion-based.

Talend

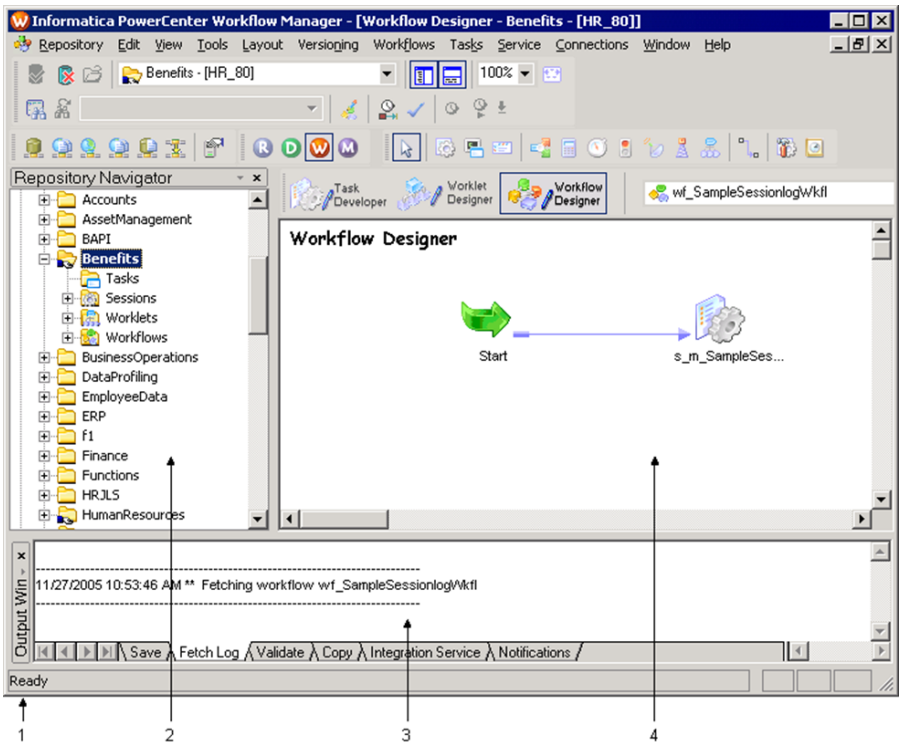
One of the more popular ETL tools in this category is Talend. Talend is an open-source data integration tool and can be used to monitor, capture, transform, and deliver the data used for load applications. Talend can connect to most of the databases like SQL, NoSQL, Oracle, and others. It can also be deployed on the Cloud, which makes it a hybrid tool, as not many ETL tools can be deployed on the Cloud. The open-source version of Talend is data integration Talend Open Studio (TOS). Data transformation looks as shown in below diagram in talend.



The ETL jobs developed in TOS can be exported in Talend Data Integration, a paid version with enhanced job monitoring and scheduling features. The ETL jobs developed in the Open Studio version can be converted either into batch processes or real-time processes using the event processing feature. It uses Talend Data Management Platform on the backend, which is highly scalable. It supports data quality tools. The tool provides a wide range of components for batch processes. The web service consumers can directly use REST calls without using the SOAP web services which makes it easier. It provides scheduling and monitoring features for the jobs in paid version using Talend Enterprise Scheduler. The development of jobs in full GUI-based tool makes the understanding easier when compared to coding in other java-based ETL tools [5].

Informatica

Informatica PowerCenter is an ETL tool used for Data Warehousing Solutions. Informatic data integration tool is highly scalable & dependable ETL tool using high performance & productivity. It is a solution to powerful & popular data warehousing & business intelligence application [1]. Informatica PowerCenter is a robust ETL tool that provides analytical solutions used widely. It is a leading data integration tool used widely around the globe. Movable data integration platform is designed to make multiple data sources joined as one view to analytics tools & applications for reporting. Informatica ETL tool extracts data from multiple heterogeneous sources of data, transform & integrate the data as business perspectives according to organization needs & data gets loaded to the target data warehouse system. It helps to create a view of all business data so that it would become a foundation for perceptive wealth of information [5].



1. Status bar
2. Navigator
3. Output
4. Main

Informatica PowerCenter is an enterprise-class data integration server that manages & costs various design, execution & administration tasks associated with the ETL process, Informatica PowerCenter is a data integration server that interacts with multiple data source systems: databases, XML files, flat files, applications focus systems & figure data warehouse systems. In ETL process, Informatica ETL Tool PowerCenter, the key tasks are Extraction, transformation & Loading (ETL). The above figure illustrates the features available in PowerCenter tool. These tasks could be done using drag & drop graphical user interface. Informatica

PowerCenter Tool needs a central repository (Informatic PowerCenter Repository) that is a set of database schema. The repository stores all metadata related to the ETL Process. The repository is designed to maintain both business & technical metadata. Business metadata defined the business definitions context & other information maintained by the user. Informatica ETL Tool PowerCenter is maintainable with Metadata Transporter Utility.

Other ETL Tools

Apache Nifi - It is an open source ETL tool created by the Apache software foundation in 2014, which can be used to process and distribute data between data sources and systems. Using the tool users can create a data flow pipeline. Nifi can be deployed on the user's own servers, allowing for data privacy, availability and control. It also allows for load balancing, backup and recovery and operating within the user's firewall [1].

Oracle Data Integrator - It is an ETL tool provided by Oracle corporation, which allows for transformations to be coded in "Java or SQL." Gravity is its primary ETL engine. It operates in two modes: an ETL mode, where data is parsed and processed before being inserted into the target warehouse, and an ELT mode, where the warehouse is transformed using SQL statements applied directly to it [4].

AWS Glue - A fully managed ETL tool provided by Amazon Web Services. It can automate a user's data extraction process and summarize it for analysis. In its base form, Glue can work where data is stored on AWS services and can run python or spark jobs. Like DataFlow, Glue has a philosophical focus on using metadata that is automated, and each connection or transformation is governed by a crawler or a. glue script created by using the console to create a pipeline.

Performance Evaluation

All ETL tools are tested crosswise on data with a predetermined length of 10,000 records before documenting the performance evaluation results. Data will be evaluated on three platforms: an Intel Core i5 processor machine with 4GB and 8GB RAM, and an Intel Core i7 processor machine with 16GB RAM and above operating windows and MYSQL. The machines will be set up on the same local Wi-Fi network with a data setting of 10,000 registered accounts with available user names, email accounts, dates of birth, passwords, and other relevant information. Memory analysis will be done to determine the server's capabilities and bounds, and 10 files will be created for input on edc and swap testing. The ETL tools will then be implemented for performance evaluation. To ensure reproducibility, information regarding the platforms, data setup, and tool implementations will be recorded.

Informatica is a market-leading ETL tool in the United States and mostly in the early domains, whereas Talend is widely used in Europe and South Asia. So a performance benchmark is established for these two ETL tools using compression, decrease, jaccard, union, and intersection as the probable factor to look for. The basis of ETL tools performance is based on common operations carried on in regular transformation. Five trial runs are carried out to mitigate randomness in time because time can be affected by various external environments. The results are successfully analyzed, examined, and found by using bar charts and tables as pictographs in a comparative manner. It is observed that Talend execution time is overall faster than Informatica. In the case of a growing number of records, the gap of performance minimizes and permits improved throughput ratio [4].

Speed and Efficiency

While speed and efficiency may seem synonymous, they can have distinct meanings. Therefore, the focus of this subsection is on speed and efficiency, specifically looking at the performance of the ETL tools with regards to speed, as well as their efficiency. To gauge the swiftness of the tools in practical usage, a group of datasets were used to execute the same dummy job on each of the tools. Each job involves reading a datafile from disk and inserting records into a database table, with the database server located on a different physical machine. During the jobs' execution, the processing times were logged to obtain the insertion time alone, following this query structure: "INSERT INTO TableName (field1, field2, …) VALUES (value1, value2, …)" [4]. The results are presented in Table 6, with the time taken in seconds, and with respect to the size in MB.

Alongside speed, efficiency is evaluated by analyzing the CPU and memory consumption of each of the tools while they are running the same job and processing the same data. This enables determining how much processing power and RAM the ETL tools require to fulfill the jobs. At each ten-second interval, the total CPU

and memory usage of the ETL tool was logged to a CSV file. The recorded values from the CSV files were processed and the average percentages are displayed in Table 7.

Scalability

This subsection focuses on the scalability of ETL tools and the ease at which the end user can scale them to cater for an increase in users or data size. Talend Open Studio for Data Integration is designed to remain performance-enduring with an increase of users or data set size. This is made possible by using multiple parallel orders for extractions and extra high-performance elementary transformation operator such as sorting and hashing. Moreover, transforming rows to key/value pairs for ETL processes can enhance performance by reducing transmission consequences [1]. Informatica Power Center is scalable due to its rich partitioning features for sources, various types of transformation and targets and by using a parallelism for data flow process. This can be a challenge, as data can get lost while the user can spare a rollback point if the job fails. Also, in contrast to Talend, Informatica requires tuning at various levels to achieve enhancement performance. For example, there are multiple configuration files in Informatica, the user must configure the grids for informatica services, the "max session task" for warehouse; this value can be higher than the "max concurrent sessions" parameter [4]. These results in Talend applicability to larger databases with no earlier tuning and minimal resources, while Informatica solution can increase requirements exponentially in hardware. The Pentaho Data Integration (PDI), also known as Kettle, is open-source ETL tool provided by Pentaho. PDI is designed to extract, transform and load processes by a graphical user interface. PDI can be a scalable tool according to the needs, as instead of the community version, there is an enterprise edition which includes more functionalities. However, in the community version has no option for clustering and partitioning therefore not easily scalable in businesses growth. However, open-source ETL tools can be installed in a Hadoop environment providing scalability to the processes, they would be able to handle more data and perform tasks more quickly after a configuration.

Conclusion

The comparative analysis of ETL tools presented in this paper highlights the diverse landscape of data integration technologies available today. The evaluation covered various aspects, including performance, scalability, ease of use, cost-effectiveness, data transformation capabilities, support for diverse data sources, integration with big data technologies, and automation features. Each ETL tool has its unique strengths and weaknesses, making it crucial for organizations to carefully assess their specific needs and operational requirements before selecting a solution.

Open-source ETL tools, such as Apache NiFi and Talend Open Studio, offer flexibility and cost advantages, making them suitable for organizations with skilled technical teams and limited budgets. These tools provide robust data transformation capabilities and extensive support for various data sources. However, they may require more effort in terms of setup and maintenance.

Commercial ETL tools, like Informatica PowerCenter and Microsoft SQL Server Integration Services (SSIS), typically offer superior ease of use, advanced features, and comprehensive support. They are well-suited for enterprises seeking streamlined data integration processes with minimal manual intervention. The higher cost of these tools is often justified by the enhanced support and reduced time-to-value they provide.

Cloud-based ETL solutions, such as AWS Glue and Google Cloud Dataflow, leverage the scalability and flexibility of cloud platforms. These tools are ideal for organizations looking to integrate and transform large volumes of data in real-time, benefiting from the seamless integration with other cloud services and the pay-as-you-go pricing model.

The choice of an ETL tool should align with the organization's data strategy, technical expertise, and budget constraints. While some tools excel in handling complex transformations and large datasets, others are designed for simpler, more straightforward ETL processes. Ultimately, the selection process should involve a thorough assessment of the tools' capabilities against the organization's current and future data integration needs.

In conclusion, there is no one-size-fits-all solution in the ETL landscape. By understanding the strengths and limitations of each tool, organizations can make informed decisions that enhance their data integration processes, driving better data management and analytics outcomes. This comparative analysis serves as a guide

to navigate the diverse ETL tool ecosystem, ensuring that the chosen solution aligns with the organization's goals and delivers maximum value.

References

[1]. J. Sreemathy, R. Brindha, M. Selva Nagalakshmi, N. Suvekha, N. Karthick Ragul, and M. Praveennandha, "Overview of ETL Tools and Talend-Data Integration," IEEE Xplore, Mar. 01, 2021.
[2]. V. Theodorou, "Automating User-Centered Design of Data-Intensive Processes," 2017.
[3]. M. S. Jamaluddin and N. F. M. Azmi, "Extraction transformation load (ETL) solution for data integration: a case study of rubber import and export information," 2016.
[4]. G. V. Machado, Ítalo Cunha, A. C. M. Pereira, and L. B. Oliveira, "DOD-ETL: Distributed On-Demand ETL for Near Real-Time Business Intelligence," 2019.
[5]. T. Čufer, "Simplification of ETL processes using Talend Platform," 2015.