Comparative Analysis of ETL Tools in Big Data Analytics

Asma Qaiser[1], Muhammad Umer Farooq[2], Syed Muhammad Nabeel Mustafa[2] and Nazia Abrar[3]

[1]Department of Computer Science, Iqra University, Karachi 76400, Pakistan
[2]Computer Science & Information Technology Dept., NED University of Engineering & Technology, Karachi 76400, Pakistan
[3]College of Computing & Information Sciences, Karachi Institute of Economics & Technology, Karachi 76400, Pakistan

Corresponding author: Asma Qaiser (email: asma.qaiser@iqra.edu.pk).

Abstract- **Massive data volumes are being created across the globe at the moment. Most social networking sites, e-commerce websites, private and public domains, healthcare systems, cloud networks, and numerous systems generate massive amounts of data. Information obtained from numerous locations might be ordered or unstructured. ETL (Extract, Transform, Load) is critical to the rising demand for faster business decisions in many modern systems. Due to the volume and speed of data, real-time ETL is built on the foundation of multi-source, unstructured data stream extraction and transformation employing disc data in dispersed environments. While processing, the complete technique is pipelined so that the result data might give some important and relevant findings to work upon. Some analytical discoveries are once again useful in decision-making. However, the analytical may differ in various statistics, graphs, and figures in many cases. It is the outcome of applying some impractical technologies for large data processing. Choosing a decent ETL tool is tough because so many are accessible. This article explores numerous tools for doing ETL processes. Our goal is to study the benefits and weaknesses of ETL Components and compare them. It will assist researchers and industry specialists in selecting the appropriate tools for their requirements.**

*Index Terms-- **Big Data Analytics, Big Data, ETL, Comparative Analysis***

I. INTRODUCTION

People generate a vast volume of data daily due to the rapid expansion of information technology. As a result, dealing with such massive amounts of data and uncovering hidden patterns and insights to make useful decisions has become a problem for all companies. To address this issue, a data warehouse is necessary, which stores large amounts of data from many sources, and this sort of data is referred to as Big Data [1].

The Extract, Transform, and Load (ETL) method is the most typical method for collecting data from many sources and integrating it into a single source of information, i.e., a data warehouse [2]. During the ETL process, data is extracted from such a resource as a relational database, XML files, flat files, or spreadsheets, transformed to data warehouse standards, and then loaded into the warehouse [3]. Figure 1 depicts the fundamental architecture of the ETL process. Although ETL is required for data warehousing and analytics, not all ETL software packages are created equal. The best ETL tool may vary depending on the circumstances and use cases [4]. This document compares the most common ETL 2022 tools [5].

This article includes an overview of ETL tools, ETL categories, and a comparative study of ETL tools. Section II contains the remainder of the paper. Section III examines the functioning of ETL tools (see Fig. 1), Section IV compares features, and Section V concludes the paper.
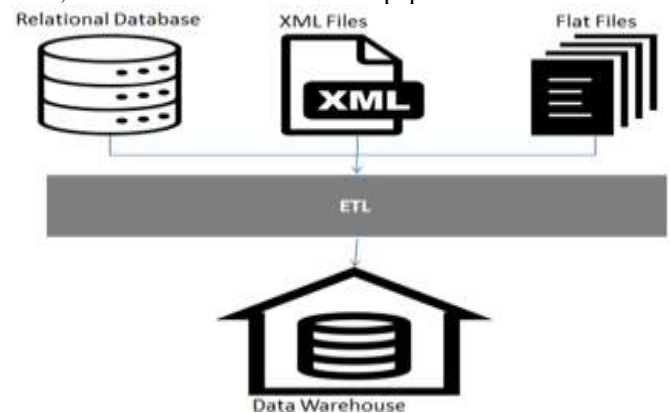

FIGURE 1.   ETL process fundamental architecture

II. ETL TOOLS CATEGORIES

ETL technologies help businesses make their data accessible relevant, and useable across disparate data systems. There are several options accessible when choosing the correct ETL tool [6]. This section identified the categories and their common characteristics, which will make it easier to select the best tool for

your needs. Most tools fall into various categories due to their powerful features [7].

A. BATCH PROCESSING
The traditional method of processing is to gather data before processing in batches. The file is acquired, analyzed, legitimated, cleaned, calculated, and aggregated before being sent to a system for additional analysis. Modern businesses cannot wait for data to be collected before processing it; they need to process data right now. e.g., IBM DataStage, Informatica, Oracle Data Integrator, InfoSphere and SSIS [8].

B. CODE-BASED/ENGINE-BASED
Programs not developed in a proprietary language are compiled and used as code-based ETL solutions. ETL programs are usually generated by code written in general-purpose languages. These solutions are typically proprietary and have performance-optimized data engines. Overall, engine-based products perform better. Equally, engine-based and code-based ETL solutions perform well. However, engine-based ETL tools are better if one does not have the skills to adapt the code. Oracle Warehouse Builder and SQL Server Integration Services (SSIS) are examples of engine-based tools [9].

C. CLOUD-BASED
Streaming data processing, real-time, scalability and connections with an ever-growing number of data sources are all advantages of cloud-based ETL technologies. Alooma, Blendo, Matilian, Stitch, and, Fivetran are examples of cloud-based ETL tools [10].

D. OPEN SOURCE
Certain people favour open-source alternatives to commercial solutions because they are less expensive. Due to their less complicated data and reporting requirements, mid-market firms and system integrators choose to support and utilize open-source ETL solutions. Open-source ETL tools include Talend Open Studio Apache Kafka, Apache NiFi, and Apache AirFlow [11].

E. GUI-BASED
Due to the application's simple user interface, Graphic User Interface (GUI) based ETL applications are the most extensively used products on the market. Drag and drop functionality is included in GUI-based data loading and analysis tools. Some GUI-based ETL tools include Pentaho, Informatica, DataStage and Abinitio.

F. REAL-TIME
Real-time data processing is required as data, and its sources evolve. The requirement for real-time data has altered ETL design. Striim, Confluent, StreamSets, and Alooma are the most common tools [12].

G. NoSQL-BASED
The ETL process, which collects digital data and transforms it into understandable reports, has existed since the beginning of the digital era. The analysis becomes more difficult to do with the growth of schema-less databases. A well-known NoSQL-based database is MongoDB. It offers both paid and free/open-source versions of its ETL tools. Transporter, MongoSyphon, Panoply, Krawler, Pentaho, Talend Open Studio, and Stitch are some more solutions that can provide data integration with NoSQL databases like MongoDB [13].

III. STUDY OF ETL TOOLS FUNCTIONALITY
The comparison analysis and details for the selected ETL tools are provided in this section. Xplenty, Talend, Alooma, AWS Glue, Informatica, and Talend.

A. INFORMATICA
In 1993, a famous ETL tool, Informatica, was invented in California. Informatica is one of the best-known tools for collecting and retrieving data from multiple sources. Informatica provides many features, such as Data Management, Data Quality, Data Security and Data integration. Informatica PowerCenter is the product name used to perform ETL operations [14].
Informatica is useful in the following scenarios:
1. A business is switching from an outdated database to a new software system.
2. To create a data warehouse in a business so that data is sent from the production/data gathering system to the warehouse
3. It serves as a tool for data cleaning, identifying, modifying, or eliminating information from inaccurate or false database records.
Informatica is a little expensive. Informatica's price varies depending on the license length, number of users, or number of servers, and Informatica offers hourly licenses. Informatica supports custom transformations written in C and JAVA. The Power Center Designer client can handle JAVA code snippets.

B. TALEND
The first data integration software, called Talend, was made available in 2006. One of its main products is the Talend open studio. It is helpful for data warehousing, data migration, ETL, and business intelligence. Talend generates Java editable code.
The most open, cutting-edge, and potent data integration tool available today is Talend Open Studio. In addition, it provides data quality, an open profiler, on-demand services, and an integrated suite. Some characteristics are metadata-driven design and execution, robust execution, business modelling, graphical development, and real-time debugging [15].

C. ALOOMA
Alooma offers real-time data streaming and uses the cloud and code engines to alter data. Alooma can capture, convert, and store data from millions (or billions) of transactions across a wide range of data sources and streams. This capacity opens up a slew of new possibilities, like evaluating historical records to improve the sales process, altering prices and inventories in real-time, integrating machine learning and artificial intelligence to construct predictive models, establishing new revenue streams, and more [16].

D. AWS GLUE

Finding, preparing, and merging data for application development, machine learning, and analytics is simple by using AWS Glue, a serverless data integration solution. It provides all the tools needed for data integration, allowing users to analyze and use data in minutes instead of weeks or months. Data engineers and ETL developers may graphically create, run, and monitor ETL workflows with only a few clicks in AWS Glue Studio. Data scientists and analysts can visually improve, clean, and standardize data with AWS Glue DataBrew without writing code. Application developers can integrate and replicate data across many stores using SQL with AWS Glue Elastic Views. [17].

E. STITCH

Because Stitch is publicly accessible, it is utilized as an open-source solution and for a straightforward ETL process. This is because open-source software and this ELT data integration platform do not require complicated transformations. ETL Stitch's cloud-based platform provides vital insights and saves time managing data pipelines. It makes data pipelines more transparent and controllable. Stitch focuses on analysis rather than pipeline maintenance [18].

F. APIFY

Apify is an ETL solution that enables aspirational businesses to capitalize on the web's full potential, the biggest data source ever produced by a human. Using this software platform, 10 distinct extraction procedures for 10 different websites with 10 configurations or constraints might be created.

The features that Apify has incorporated allow users to find new customers and gather information about them, generate large-scale datasets from the web, build new products and services using data aggregation from the web, monitor competitors online, automate repetitive tasks, and generate insights into the market from publicly available data on the website [17].

G. SKYVIA

One of the most well-known cloud-based ETL tools, Skyvia offers users strong support for data migration, backup, and integration. Being a SaaS programme, all users need to access it is a web browser and stable internet connectivity. It is universal cloud data platform that allow data integration, backup management, connectivity, automated data collection, transfer of business data among cloud app automatically with just a few clicks, and share data via REST API to connect with multiple consumers in real-time [19].

Skyvia offers different pricing plans for each product that is implementable on any size of business, from small startups to enterprise companies.

H. HEVO

Hevo is an ETL tool based on no-code Data Pipeline. It offers a completely managed solution to organize data integration from hundreds of sources and let it directly load into an organization's data warehouse [20]. Significant features of Hevo include the following:
It is full management as it does not require any management or maintenance.

- It has a simple interface to modify, enrich and prepare the data according to the requirement.
- It supports real-time data migration to make data available anytime for analysis.
- It automatically detects the schema of extracted data and maps it to the destination.

I. FIVETRAN

A cloud-based ETL solution called Fivetran offers one of the most flexible integration options and high-end performance, supporting over 90+ SaaS sources in addition to other databases and other bespoke connectors [21]. It enables deploy automated ETL pipelines in a matter of minutes and is fully managed. It has a user-friendly platform with a short learning curve that enables data integration and loading into several data warehouses. Fivetran is also a cloud-based ETL tool developed for data analysts and data engineers. It comes with various built-in integrations and functionalities that help users push data into the data warehouse.

J. AGENTY

Agenty is a scalable cloud-based SaaS tool for Sentiment analysis, Price monitoring, Data scraping, OCR, Web-automation, Machine translation, Change tracking and many other functionalities and options. It helps to overcome the complexity of RPA. Its built-in API allows us to build the web automation experience ever repetitive tasks. It has much functionality, including the following:

- Allows creating an agent with just a few clicks and obtaining results instantly.
- Helps to get automatic email alerts when the job has been completed.
- Helps to view the activity logs for all runs with events and messages.
- Has the ability to process millions of pages of hundreds of customers.
- Provide an option to re-use all data processed ever for analytics.

K. HUBDOC

Hubdoc is a data extraction tool developed to assist practices simplifies document workflow by capturing and storing client bills and receipts. Hubdoc allows the accountants and bookkeepers to be offered a free Hubdoc partner organization to test the product and features [22].

l. GREPSR

Grepsr is a simple and streamlined web scraping service platform that assists in getting and consuming data [23]. Grepsr provides the management and scraping solution to the user according to the need without any complications and tool configuration.
It provides several features, including the following:

- Initial Project Consultation.

- Instrument Web Crawlers & Data Collection.
- Hassle Free Maintenance.

M. XPLENTY

Xplenty is an ETL platform that does not require any programming. Companies that require pre-built integration procedures and place a higher priority on data security make use of Xplenty. The easy data integration, processing, and preparation activities are made simple by the point-and-click interface, which saves time and effort. It offers a graphical user interface for creating data pipelines between a large number of sources and destinations. Xplenty comes with a slew of SaaS, applications. mongoDb, mysql, postgresql, amazon, redshift, Googlecloud. Xplenty's advantages include scalability, Security, and customer service. It also adheres to legislation such as HIPPA, GDPR, and CcPA to ensure consistency[24].

IV. FEATURE COMPARISON

We have identified key characteristics for a successful ETL solution based on a survey conducted by several researchers and online sources [25]. Big enterprises are looking for capabilities like 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). Table I demonstrates that, regardless of whether a feature is present in the tool, some features are either partially present or not present. L denotes the given tool's constrained feature coverage [26].

TABLE I
COMPARISON OF FEATURES

Tools	T	R	G	C	S	N	P	M	A
Informatica	Y	Y	Y	Y	L	N	Y	Y	N
Talend	Y	Y	Y	Y	N	N	Y	N	N
Alooma	Y	Y	N	Y	N	N	N	N	L
AWS Glue	Y	Y	Y	Y	N	Y	N	Y	Y
Stitch	Y	N	N	Y	N	Y	Y	N	Y
Xplenty	Y	Y	Y	Y	N	Y	N	N	L
Apify	Y	Y	Y	Y	Y	N	Y	N	Y
Skyvia	Y	Y	Y	Y	N	N	N	Y	Y
Hevo	Y	Y	Y	Y	Y	N	N	Y	Y
Fivetran	Y	Y	Y	Y	Y	Y	N	Y	Y
Agenty	Y	Y	Y	Y	N	N	Y	Y	Y
Grepsr	Y	Y	Y	N	N	N	Y	N	Y

When choosing an ETL solution, the company must first determine which characteristics are critical to its success. A company might choose an ETL solution for data integration based on the features it offers, as discussed in Table II.

TABLE II.
ETL TOOL SELECTION FOR AN ORGANIZATION

Tools	Type of Organization	Organization Size
Informatica	• Healthcare • Retail • Finance	Enterprise
Talend	• Media • Information technology and services	Enterprise
Alooma	• Education • Consumer Electronics	Mid-Market
Stitch	• Computer software • Information technology and services	Small business
Xplenty	• E-Learning • Advertising • Marketing	Market Small business Mid-
Apify	• E-commerce • Marketing • News Website • Travel site	Enterprise
Skyvia	•Information technology and services • Marketing • Advertising	Enterprise
Hevo	• Customer Support • E-commerce • Non-profit organization • Cloud-storage service	Mid-Market
Fivetran	• Salesforce • Financial Analytics • Scientific Simulation	Small business
Agenty	• E-commerce • Cloud storage service • Virtual Travel Agent •Marketing and Advertising	Mid-Market
Grepsr	• Retail • E-commerce •Marketing and Advertising	Mid-Market

The features of ETL tools are discussed in Table III.

V. CONCLUSION

This paper focuses on the various types of ETL tools and packages available. Because of their features and implementation methodologies, many tools can be classified into multiple categories. Choosing the right ETL tool for the organization is critical in every business. The selected ETL tool should not be customized more than is required. If real-time updates or the handling of stream datasets are not required, one can use any simple solution that meets the needs. ETL tools can be developed using open-source technology if one works with a huge amount of data or streaming data.

Features	ETL Tools
History and cost	• Informatica: Active since 1993. • Talend: 2016, launching year, less expensive because it is an open-source • Alooma: 2019 launching year, costly between $1000 and $15000 per month • Stitch: open-source • Aws Glue: Launch 2017, costly • Xplenty: launch 2012, Standard plans range from $100 to $1,250 per month • Apify: launched 2019, standard plan ranges from $0 to $500 • Skyvia: Launched 2014, standard plan ranges from $15 to $800 • Hevo: Launched 2016, Standard plan ranges from $0 to $1000. • Fivetran: Launched 2012, Standard plan is $1.50 Per credit • Agenty: Launched 2016, Standard plan ranges from $29 to $250 • Grepsr: Launched in 2012, Standard plan starts from $149
Component of ETL process	• Informatica: Informatica Repository Manager, PowerCenter Designer, Workflow Manager, Workflow Monitor and Administrator Console • Talend: tMysqlConnection, tMysqlInput, tMysqlOutput , tFileInputDelimited and more • Alooma: ETL users on Google Cloud Platform can focus more on the project and less on the technical side of things by leveraging its automation capabilities. • AWS Glue: Data Catalog, ETL Engine etc.
Supported Language	• Informatica: C, JAVA • Talend: JAVA • Stitch: No coding • Xplenty: no coding • Alooma: Python • Aws Glue: Python, Scala • Apify: Python, Rust • Skyvia: No coding • Hevo: No coding, Use JSON APIs • Fivetran: requires SQL coding skills • Agenty: No Coding • Grepsr: No Coding
Performance Optimization	• Informatica, The goal of performance tuning is to optimize session performance by eliminating performance bottlenecks. We can improve it by tuning all the bottlenecks, e.g., Optimizing Power Center components. • Xplenty: optimize performance with AI/ML- driven operational monitoring • Turbocharging Maps by Talend. Speed up Talend ETL Operations or Else • Alooma: they gather clickstream data from the busy websites of their clients. • Aws Glue: With the help of the file
	• grouping capability in AWS Glue, numerous files can be combined into one Spark task. • Apify: permits different web services and APIs to be connected, allowing data to flow between them. step up your unique data processing and computing. • Skyvia: offers contemporary cloud agility and scalability, eliminating need of manual upgrades. • Hevo: helps to get data pipelines up and running in a few minutes • Fivetran: automatically and continually update, removing attention from ETL to concentrate on game-changing insights. • Agenty: allows to track competitor price, looking for high-quality data feed to train any AI algorithm, has built-in API for web scraping on cloud. • Grepsr: allows quick turnaround times with the highest quality standards without programming skills or complicated tool understanding.
Metadata Management	• Informatica: Informatica Metadata Manager is a web-based metadata management tool that can browse and analyze metadata from disparate metadata repositories. • Talend: Talend Metadata Manager provides an automated way to parse and load different types of metadata. • Alooma: SAS provides a Data Management platform consisting of more than twenty ... Security and compliance, data governance, and metadata management • Aws Glue: AWS Glue uses the AWS Glue Data Catalog to store metadata about data sources, transforms, and targets.
Real time Refreshment	• Informatica: can be used PowerCenter to process data in real-time. • Talend: can refresh it any time by clicking Refresh on the toolbar • Aws Glue: use AWS Glue Elastic Views to integrate and continuously replicate data across various data stores in near-real time • Alooma: emphasizes the value of a real-time data pipeline service that allows for multi-automation capabilities. • Aws Glue: Data Catalog, ETL Engine etc.

According to a study, numerous critical parameters exist for choosing ETL tools. It is divided into categories based on the type of organization, the intended purpose, current data storage, connectivity, real-time processing, data integration choices, visualization assistance, and cost. There are many different ETL tools on the market. However, there is always room for innovation, particularly in open-source technology for small enterprises, to give more functionality in a single tool at a lower cost.

CONFLICT OF INTEREST

The Authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] F. d. A. Vilela and R. R. Ciferri, "A survey of real-time etl process applied to data warehousing environments," in *ITNG 2022 19th International Conference on Information Technology-New Generations*. Springer, 2022, pp. 91–96.

[2] A. Dhaouadi, K. Bousselmi, M. M. Gammoudi, S. Monnet, and S. Hammoudi, "Data warehousing process modeling from classical approaches to new trends: Main features and comparisons," *Data*, vol. 7, no. 8, p. 113, 2022.

[3] J. Y. A. Moura and B. Z. Cadersaib, "Effort estimation method for extract transfer load (etl) big data projects," in *2022 2nd International Conference on Information Technology and Education (ICIT&E)*. IEEE, 2022, pp. 160–167.

[4] E. Mehmood and T. Anees, "Distributed real-time etl architecture for unstructured big data," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3419–3445, 2022.

[5] B. Oliveira, M. Leite, O. Oliveira, and O. Belo, "A service-oriented fsssrame-' work for etl implementation," in *EPIA Conference on Artificial Intelligence*. Springer, 2022, pp. 636–647.

[6] J. C. Nwokeji and R. Matovu, "A systematic literature review on big data extraction, transformation and loading (etl)," *Intelligent Computing*, pp. 308–324, 2021.

[7] J. Sreemathy, R. Brindha, M. S. Nagalakshmi, N. Suvekha, N. K. Ragul, and M. Praveennandha, "Overview of etl tools and talend-data integration," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2021, pp. 1650–1654.

[8] M. Patel and D. B. Patel, "Progressive growth of etl tools: A literature review of past to equip future," *Rising Threats in Expert Applications and Solutions*, pp. 389–398, 2021.

[9] N. Biswas, A. Sarkar, and K. C. Mondal, "Efficient incremental loading in etl processing for real-time data integration," *Innovations in Systems and Software Engineering*, vol. 16, no. 1, pp. 53–61, 2020.

[10] A. Erraji, A. Maizate, and M. Ouzzif, "New etl process for a smart approach of data migration from relational system to mongodb system," in International Conference on Digital Technologies and Applications. Springer,2022, pp. 131–140

[11] R. Aluvalu and M. Jabbar, "Handling data analytics on unstructured data using mongodb," 2018.

[12] R. C. L'Esteve, "Mapping data flows for data warehouse etl," in *The Definitive Guide to Azure Data Engineering*. Springer, 2021, pp. 215–236.

[13] H. Homayouni, "Testing extract-transform-load process in data warehouse systems," in *2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE, 2018, pp. 158–161.

[14] L. Ehrlinger and W. W¨oß, "A survey of data quality measurement and monitoring tools," Frontiers in Big Data, p. 28, 2022.

[15] D. Y. Siregar, H. Akbar, I. B. P. A. Pranidhana, A. N. Hidayanto, and Y. Ruldeviyani, "The importance of data quality to reinforce covid-19 vaccination scheduling system: Study case of jakarta, indonesia," in 2022 2nd International Conference on Information Technology and Education (ICIT&E). IEEE, 2022, pp. 262–268.

[16] S. Jankovi'c, S. Mladenovi'c, D. Mladenovi'c, S. Veskovi'c, and D. Glavi'c, "Schema on read modeling approach as a basis of big data analytics integration in eis," Enterprise Information Systems, vol. 12, no. 8-9, pp. 1180–1201, 2018.

[17] G. Stefanov, "Analysis of cloud based etl in the era of iot and big data," in Proceedings of International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE). International Conference on Application of Information and Communication ..., 2019, pp. 198–202.

[18] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. De-la Hoz-Franco, and E. De-La-Hoz-Valdiris, "Trends and future perspective challenges in big data," in *Advances in intelligent data analysis and applications*. Springer, 2022, pp. 309–325.

[19] F. Aqlan, J. C. Nwokeji, and A. Shamsan, "Teaching an introductory data analytics course using microsoft access® and excel®," in *2020 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2020, pp. 1–10.

[20] [Online] Leverage data effortlessly. Available online: https://hevodata.com/ (Accessed on: Nov 25, 2022)

[21] [Online] The most reliable data pipelines you'll never build. Available online: https://www.fivetran.com/ (Accessed on: Nov 26, 2022)

[22] [Online] All your bills and receipts in one place: Available online: https://www.hubdoc.com/ (Accessed on: Nov 27, 2022)

[23] [Online] Managed data acquisition, perfected at scale: Avaliable online: https://www.grepsr.com/ (Accessed on Nov 28,2022)

[24] [Online] Turn Your Data Warehouse into a Data Platform: Avaliable online: https://www.integrate.io/ (Accessed on Nov 29, 2022)

[25] S. Bimonte, E. Gallinucci, P. Marcel, and S. Rizzi, "Data variety, come as you are in multi-model data warehouses," *Information Systems*, vol. 104, p. 101734, 2022.

[26] J. Sreemathy, K. N. Durai, E. L. Priya, R. Deebika, K. Suganthi, and P. Aisshwarya, "Data integration and etl: A theoretical perspective," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2021, pp. 1655–1660.