An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing

1

AB STRAC T
The extraction, transformation, and loading (ETL) process is a crucial and intricate area of study that lies deep within the broad field of data warehousing. This specific, yet crucial, aspect of data management fills the knowledge gap between unprocessed data and useful insights. Starting with basic information unique to this complex field, this study thoroughly examines the many issues that practitioners encounter. These issues include the complexities of ETL procedures, the rigorous pursuit of data quality, and the increasing amounts and variety of data sources present in the modern data environment. The study examines ETL methods, resources, and the crucial standards that guide their assessment in the midst of this investigation. These components form the foundation of data warehousing and act as a safety net to guarantee the dependability, accuracy, and usefulness of data assets. This publication takes on the function of a useful guide for academics, professionals, and students, despite the fact that it does not give empirical data. It gives students a thorough grasp of the ETL paradigm in the context of data warehousing and equips them with the necessary skills to negotiate the complex world of data management. This program equips people to lead effective data warehousing initiatives, promoting a culture of informed decision-making and data-driven excellence in a world where data-driven decision-making is becoming more and more important.

KEYWORDS
ETL; ETL process; ETL techniques; ETL tools; ETL evaluation

1 Introduction

A Data Warehouse (DW) is described as a “collection of integrated, subject-oriented databases designated to support the decision making process” [1] by providing customers with exclusive access to several sources, it seeks to enhance decision making. It aims to improve decision making by giving users unique access to several sources. Data warehouses serve as a single store for historical and current data from diverse sources. It mostly comprises of historical data derived from current data and transaction data obtained from a variety of sources [2]. It is frequently differentiated by a collection of linked, subject-oriented, volatile, and time-variant databases. In order for our DW to accomplish its purpose of facilitating business analysis, we must load it on a regular basis. To do this, data from one or more operating systems must be obtained and put into the DW. The objective in DW circumstances is to

reorganize, aggregate, and construct huge volumes of data across several systems, resulting in a new collective knowledge base for business intelligence [3]. The Extraction-Transformation-Loading (ETL) process lies at the heart of DWs, which integrate data from multiple sources into single source. The ETL was basically proposed as data integration and loading method for analysis and computation. As databases became more widespread in the 1970s, they finally became the primary mechanism for data processing in DW operations [4,5].

ETL cleanses and organizes data using a set of business rules to meet specific business intelligence requirements, such as monthly reporting, but it may also handle more complicated analytics to improve back-end operations or end user experiences. An organization will frequently employ ETL to:

- Extractdatafromlegacysystems
- Cleansethedatatoimprovedataqualityandestablishconsistency
- Loaddataintoatargetdatabase

This article offers a thorough explanation of the ETL process, which makes important contribu-tions to the field of data warehousing. It spreads fundamental information about data warehousing and the particular difficulties it presents, especially with regard to quality control, data integration, and managing a variety of data sources. Additionally, the article provides readers with a comprehensive overview of this important technology environment by extensively exploring ETL processes, tools, and their assessment criteria. The paper provides practitioners, researchers, and students with a valuable practical guide, even in the absence of empirical results. This empowers them to carry out data warehousing projects successfully and make informed decisions in an increasingly data-driven environment, thereby improving business intelligence and decision-making processes.

Relevant statistics that clearly show the difficulties and ramifications highlight the need of ETL operations in data warehousing. First off, the data sphere is expected to increase exponentially, with 175 zettabytes of data estimated globally by 2025 (IDC), underscoring the enormous problem of data management. Second, according to Gartner, organizations that have poor data quality suffer a significant financial loss of $15 million on average each year. This demonstrates the vital part ETL plays in improving the quality and cleaning of data. Thirdly, there is a lot of complexity in data integration. Informatica revealed that 42% of organizations struggle with data integration and migration, highlighting the difficulties in combining data from several sources into a single repository. Finally, Dresner Advisory Services notes that 97% of organizations consider business intelligence and analytics to be essential to their operations, demonstrating the broad acceptance of these technologies and the importance of efficient ETL procedures in enabling these data-driven efforts. These statistics are presented in Table 1.

Table 1: Key statistics illustrating the significance of ETL processes in data warehousing

Statistics	Datagrowthrate	Dataqualityimpact	Dataintegration complexity	Businessintelligence adoption
Values	175ZBby2025 (IDC)	$15Mimpact/yr (Gartner)	42%strugglewith DataIntegration (Informatica)	97%considercrucial Tooperations (Dresner)

This study focuses on describing the ETL tools, types of tools, tools evaluation, ETL testing, and testing techniques and challenges. The remaining study is also organized in this regard as Section 2

discusses the ETL tools with its working mechanism. Section 3 illustrates the ETL testing with its challenges and comparison of testing techniques. Section 4 presents the tools types and details with its evaluation criteria, and Finally, Section 5 concludes this study.

2 How ETL Tool Works

There are 3 steps that are involved in the ETL process shown in Fig. 1 [6,7].

Figure 1: ETL steps

2.1 Extraction

The structured or unstructured data is retrieved from its source and consolidated into a single repository in this process. ETL solutions automate the extraction process and produce a more efficient and dependable workflow when dealing with big amounts of data from various sources [8].

During this step of ETL architecture, data is taken from the source system and placed in the staging area. Transformations, if any, are carried out in the staging area to ensure that the source system's performance is not jeopardized. If incorrect data is transported directly from the source into the DW database, rollback will be tough. The staging area allows you to inspect extracted data before moving it to the data warehouse [9].

The DM must integrate systems that use different DBMS, hardware, operating systems, and communication protocols. Legacy programmed such as mainframes, bespoke applications, point-of-contact devices such as ATMs and call switches, text files, spreadsheets, ERP, data from suppliers and partners, and so on are examples of sources [10,11].

As a result, a logical data map is necessary before data can be accessed and physically loaded. This data map shows the relationship between source and destination data.

Three Data Extraction methods:

- FullExtraction
- PartialExtraction-withoutupdatenotification
- PartialExtraction-withupdatenotification

Extraction, regardless of method, should have no effect on the performance or response time of the source systems. These are production databases that are updated in real time. Any slowness or locking might harm the company's profit line.

Some validations are done during Extraction:

- Reconcilerecordswiththesourcedata
- Make sure that no spam/unwanted data loaded
- Datatypecheck
- Removealltypesofduplicate/fragmenteddata
- Checkwhetherallthekeysareinplaceornot

2.2 Transformation

To increase data integrity, the data must be changed, which includes sorting, standardizing, and removing redundant data. This process guarantees that raw data arriving at its new location is entirely compliant and ready for usage. In its raw form, the data acquired from the source server is useless. As a result, it has to be cleaned, mapped, and modified. In actuality, this is the important stage when the ETL process adds value and transforms data to generate intelligent BI reports. It is a fundamental ETL concept that involves applying a sequence of functions on extracted data. Data that does not require transformation is known as direct move or pass through data [12].

During the transformation process, you may perform particular actions on data. Assume the user needs sum-of-sales revenue that does not exist in the database. Alternatively, if the initial and last names in a table are in separate columns, they may be combined before loading [13].

Following are data integrity problems:

- It is conceivable that different applications generate different account numbers for the same
- consumer.
- Other ways to identify a company name, such as Google or Google Inc.
- The usage of distinguishing names such as Cleaveland or Cleveland.
- Some critical data files have been left blank.
- Variations on the same person's name, such as Jon, John, and so on.
  Invalid items gathered at the POS as a result of human error.

Validations are performed during this phase:

- Filtering-Retrieve only particular columns.
- Data normalization with rules and lookup tables.
- Handling Character Set Conversion and Encoding.
- Units of Measurement Conversions such as Date Time Conversion, financial conversions, numerical conversions, and so on.
- Limiting Data Values-For instance, ensuring that the age does not exceed two digits.
- Verifying Data Transfer from the Staging Area to the Intermediate Tables.
- Required information should not be left blank.
- Vacuuming (for example, mapping NULL to 0 or Male to "M" and Female to "F" and so forth).
- Separating a column into several columns and merging multiple columns into a single column.
- Moving rows and columns.
- Combine data by using lookups.
- Using any sophisticated data validation (e.g., if the first two columns in a row are empty then it automatically rejects the row from processing).

2.3 Loading

The data is imported into the final destination during this phase of the ETL procedure (data lake or data warehouse). The data can be imported in bulk or at predefined intervals (incremental load). The final step in the ETL process involves transferring data into the main data warehouse database. Due to the large amount of data that needs to be processed in a limited amount of time, typically overnight, it is crucial that the data loading process is optimized for efficiency and speed [12].

When a load failure occurs, it is important to have recovery mechanisms in place that allow operations to resume from the point of interruption without putting the integrity of the data at risk.

The administrators of the data warehouse must be responsible for keeping track of the loading process, ensuring its continuation or termination based on the performance of the server.

2.4.1 Types of Loading

- Initial Load: it consists of filling all DW tables.
- Incremental Load: implementing continuous adjustments as needed on a regular basis.
- Full Refresh: deletes the contents of one or more tables and reloads them with new data.

2.4.2 Load Verification

- Check that the key field data is not missing or null.
- Run modeling views against the target tables.
- Verify that the combined numbers and derived metrics are correct.
- Data checks in the dimension and history tables.
- Examine the BI reports generated by the loaded fact and dimension tables.

3 ETL Testing Techniques

Before we begin the testing process, we must first establish the suitable ETL Testing approach. To ensure that the right approach for applying ETL testing is chosen, we must obtain consent from all team members [3]. Fig. 2 depicts many sorts of testing procedures that may be utilized.

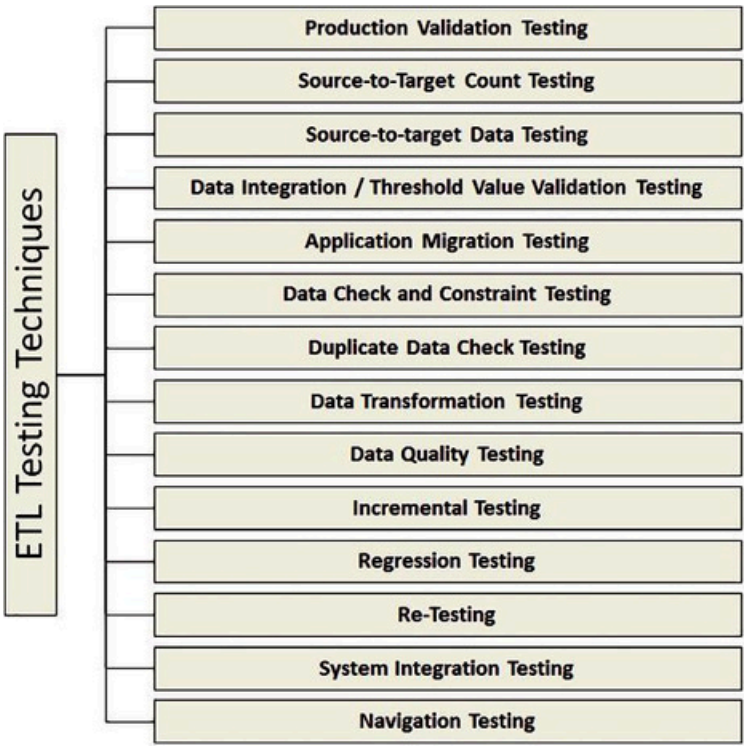

Figure 2: ETL testing techniques

Production Validation Testing (PVT): If we want to do analytical reporting and analysis, we must preserve data correctness. Data that has been transferred to the production system is subjected to validation testing. It entails assessing system information and comparing it to source data [14].

Source-to-Target Data Testing (STDT): Testers can use this testing method to verify the accuracy of data transfer from the source to the target system. This involves checking the validity of the data after it has undergone transformation, both in the source system and the destination system it is connected to. This type of examination can be time-consuming and is commonly used in industries such as medical, academic, and financial [15].

Source-to-Target Count Testing (STCT): In situations where time is limited, we can use the source-to-target count strategy to expedite the testing process. This involves determining the number of records in both the source and destination systems, without examining the actual data in the target system. This strategy is not effective if the data has been sorted in ascending or descending order after mapping [16].

Application Migration Testing (AMT): Application migration testing is automatically performed when we compare data from a previous application to data from a current application. We use this to see if the information gathered from prior applications matches that of the current application system. This testing saves a substantial amount of time [17].

Data Integration/Threshold Value Validation Testing (DI/TVVT): In this testing, the tester analyses data series. Every origin component in the target system is examined to determine if the values correspond to the predicted results. It entails integrating data from many source systems into the target system after transformation and loading [18].

Data Check and Constraint Testing (DCCT): Several constraints checks are performed in this testing, including data length, data type, and index check. This Tester is in charge of the following duties: NULL, NOT NULL, and UNIQUE are all examples of foreign keys [19].

Duplicate Data Check Testing (DDCT): This testing method checks for the presence of duplicate data in the target system. When a large amount of data is present in the target system, the risk of duplicated data in the production system increases, which can negatively impact the accuracy of analytical testing results [20].

Data Transformation Testing (DTT): This testing approach is time-consuming as it involves executing multiple SQL queries for each individual record to verify the accuracy of the transformation rules. The tester must first run these SQL queries on each record before comparing the results to the data in the target system [21].

Incremental Testing (InT): This testing method can be applied if the insertion, deletion, and update commands are executed in the correct order to achieve the desired outcome. This testing approach verifies both old and new data in a step-by-step manner [22].

Data Quality Testing (DQT): This testing encompasses a variety of checks, including date verification, numerical accuracy checks, checks for missing data (null checks), and other relevant tests. To address issues with incorrect characters, such as incorrect capitalization, the tester may use SYNTAX TEST. If the data needs to align with a specific model, the Reference Test may be employed to ensure consistency [18].

Regression Testing (ReT): It is utilized to provide additional functionality that allows testers to look for new errors, after which we can make changes to data transformation and aggregation techniques. Regression refers to data errors that arise during regression testing [23].

System Integration Testing (SIT): To carefully test each component of a system before adding modules. The importance of system integration testing cannot be overstated. System integration may be classified into three types: hybrid, top down, and bottom up [22].

Retesting: When we run the test after we finish coding, we are retesting [24].

Navigation Testing (NaT): We cover all components of the front-end, report contains information in many areas, aggregates, and calculation, and so on in this testing. System front-end testing is another term for navigation testing [25].

3.1 Comparison of ETL Testing Techniques

The testing approaches are contrasted in terms of duration, testing objectives, and data type. The comparison demonstrates which approach is faster, better, and simpler for ETL testing [3]. Table 2 presents the comparison of ETL testing techniques.

Table 2: Comparison of ETL testing techniques

S. No.	Technique	Time duration	Testingobjective	Type of data
1	PVT [14]	Dependon production system	Testingofdataregarding productionsystem	Source data
2	STCT [16]	Timesaving	Conductingtesting procedures when there is limited time available.	Source and target data
3	STDT [15]	Time consuming	Carryingoutprojectsinboth themedicalandacademic fields.	Source and target data
4	DI/TVVT [18]	Time consuming Timesaving	Conductingtestingto examinetherangeofdata.	Multiple sources of data
5	AMT [17]		Automaticallyimplementing the transfer of existing application to the current one.	Data of application
6	DCCT [19]	Time consuming	Conductingtestingtoidentify andaddressduplicatedatain the target system.	Data of table
7	DDCT [20]	Dependon data Time consuming	Whenduplicateinformation existsinthetargetsystem,it can be addressed.	Data of table
8	DTT [21]		Ifacomparisonbetweenthe targetdataandtheoutcomeis necessary.	Compare the output and target data

(Continued)

Table 2 (continued)

S. No.	Technique	Time duration	Testing objective	Type of data
9	DQT [18]	Time consuming	This testing will be utilized when evaluating the data's quality.	All the data
10	InT [22]	Step by step checks	For verifying the insert, update, and delete statements.	Data of database
11	ReT [23]	Depend of data	In order to evaluate the new functionality, modifications to the transformation data will be made.	Any type of data
12	Re-Testing [24]	Depend on data	This testing will be utilized to re-evaluate your code.	Check everything
13	SIT [22]	Convenient in small system	Individual testing of the system's components.	Focus the mainly on interfaces and flow of data
14	NaT [25]	Time consuming	To check front end output	Include data in various fields

3.2 ETL Testing Challenges

Both ETL testing and database testing are distinct, and we must overcome several problems during the ETL testing procedure. Some common challenges are highlighted here:

- Datamayvanishthroughouttheetlprocedure.
- Thedatabasemaycontainincorrect,inadequate,orrepetitivedata.
- ConductingETLtestingonthetargetsystemcanbedifficultwhenthedatawarehousesystem holds real data, as the amount of information might be substantial.
- Designingandcreatingtestcasesbecomeschallengingwhenthedatasizeisextensiveand complex.
- ETLtestersareunawareoftheconsumeroutlinedesiresanddatatradeoutflow.
- ETL testing incorporates several complicated SQL concepts for data validation in the target system.
- Tofocusonmappingdata.Mostofthetime,thetestersareunawareofthesource.
- Thegreatestpossibledelayisobservedduringthecreationandtestingofaprocedure.

3.3 Common ETL Challenges

The field of ETL is constantly evolving and new software is developed to address issues that previous versions were not equipped to handle. This is due to the continuously changing market and the need to handle new data challenges such as increasingly complex data sources and the need to scale. The challenges faced in ETL are depicted in Fig. 3.

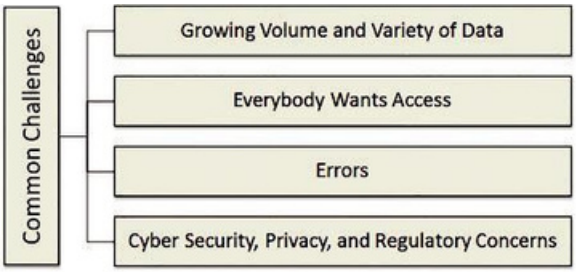

Figure 3: Common ETL challenges

Growing Volume and Variety of Data: Organizations are faced with a growing challenge as they have to manage a larger and more diverse amount of data. The increased volume and variety of data is making the design, implementation, and management of ETL pipelines more complex. The need for larger intermediate storage, as well as the costs of ETL execution and CPU resources, are also a concern. Although there have been advancements in ETL technology, especially with the rise of cloud-based solutions, there is still a fear that the growth of data may eventually surpass the ability of an organization to handle it efficiently. Thus, it is crucial for long-term success to ensure that the ETL process is scalable and can handle future growth [26].

Everybody Wants Access: As companies accumulate and manage increasing amounts of data, they face growing demand for data access, not just from data experts like scientists and analysts, but also from business users who are less familiar with data management. Everyone wants the data quickly and in a form that is easy to use. This leads to increased workload for data engineering teams who must create new data extraction and transformation processes, which in turn extends the time it takes for data scientists and analysts to derive valuable insights. It can be difficult to find the right balance between providing accurate and properly formatted data to the right people in a timely manner without compromising speed or efficiency [27].

Errors: The arrival of new data types often means new types of errors that need to be addressed. Data cleansing is crucial during the transformation stage of ETL, but the methods used for cleaning one type of data source may not be effective for another source or when loading data into a different warehouse. This can result in issues like duplicates, missing data, and other errors that can compromise the integrity of the data. It is worth noting that the more manual coding involved in the ETL process, the greater the likelihood of errors occurring, especially when using human-written Python commands instead of automated ETL tools [28].

Cyber Security, Privacy, and Regulatory Concerns: Companies need ETL solutions to ensure that their confidential information is protected and only shared with authorized individuals, whether they are facing a direct attack on their systems or complying with new regulations. Additionally, stringent data retention regulations and policies are causing businesses to need access to historical data, so they require ETL systems that can easily work with outdated data formats [29].

The real-time ETL applications provide a number of important technological problems. Unlike typical batch processing, the ETL method has particular obstacles when handling data that is created or changed in real-time circumstances. One major problem is the velocity of incoming data, as the system needs to analyses and send data quickly in order to keep accuracy and relevance. Since real-time ETL operations have to cope with continually flowing data while averting conflicts and errors, ensuring data consistency is essential. Because real-time ETL is resource-intensive, it might need a lot of processing power to keep up with the flow of data. Another difficulty is handling errors, as any

problems with the data flow have to be fixed very away to avoid data loss. To adjust to changes in the volume of data, scalability is necessary, necessitating careful design and infrastructure investment. The procedure becomes more difficult due to the variety of data sources, which include both structured and unstructured data. Sensitive data must be protected by meeting security and compliance standards, as well as by monitoring, alerting, and data quality assurance. Mitigating downtime is essential because system updates and maintenance must be done carefully to prevent hiccups in the flow of real-time data. To sum up, real-time ETL presents technological difficulties that necessitate cutting-edge instruments, technologies, and best practices to guarantee prompt, precise, and secure data processing for well-informed decision-making.

4 Types of ETL Tools

ETL tools are classified according to their infrastructure and supporting organization or vendor. Fig. 4 depicts these classifications.

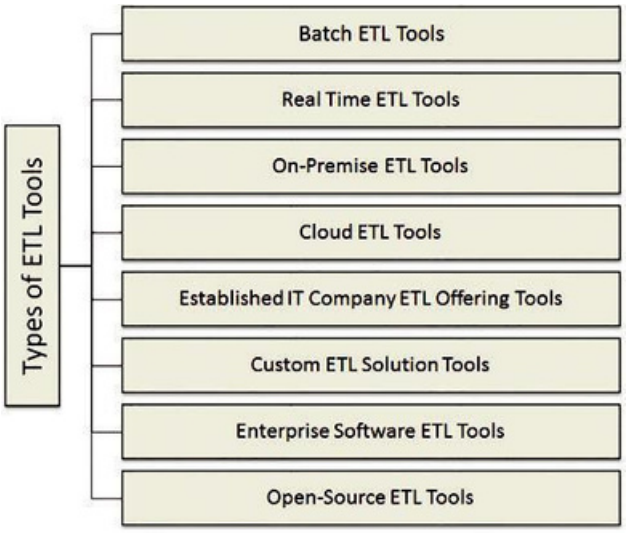

Figure 4: ETL categories

Batch ETL Tools: Batch processing is utilized to obtain data from the source systems in this sort of ETL technology. In batches of ETL operations, the data is extracted, converted, and loaded into the repository. It is a cost-effective strategy since it makes use of limited resources in a timely manner [30].

Real-Time ETL Tools: Real-time ETL solutions extract, cleanse, enhance, and load data to the destination system. These technologies provide quicker access to information and shorter time to insights. These ETL technologies are getting more popular among organizations as the necessity to acquire and analyze data in the lowest amount of time has increased [22].

On-Premise ETL Tools: Many businesses still use outdated systems that have both the data and the repository on-premise. The primary motivation for such an approach is data security. That is why businesses prefer to have an ETL tool on-site [31].

Cloud ETL Tools: As the name implies, these tools are installed on the cloud since various cloud- based apps are an important component of corporate design. To handle data transmission from these apps, businesses use cloud ETL technologies. Cloud-based ETL technologies enable enterprises to benefit from flexibility and agility during the ETL process [32].

Established Information Technology (IT) Company ETL Offering Tools: Over the years, major technology companies such as Informatics, IBM, Oracle, and Microsoft have been offering ETL (Extract, Transform, Load) solutions. Initially, these solutions were designed for batch processing on premise, but now they come with user-friendly graphical interfaces that make it easy for users to build ETL pipelines connecting different data sources. These solutions are often part of a larger platform and are attractive to organizations that need to work with older, legacy systems [33].

Custom ETL Solution Tools: Enterprises with internal data engineering and support teams can create custom tools and pipelines using SQL or Python scripting languages. Companies with adequate engineering and development skills can also take advantage of open-source options like Talend Open Studio or Pentaho Data Integration to develop, run, and optimize ETL pipelines. However, this approach offers greater customization and versatility, but also requires more administration and upkeep compared to ready-made solutions [34].

Enterprise Software ETL Tools: Commercial firms provide and support enterprise software ETL solutions. Because these firms were the first to push ETL technologies, their solutions tend to be the strongest and mature in the industry. This involves providing graphical user interfaces (GUIs) for designing ETL pipelines, support for the majority of relational and non-relational databases, as well as substantial documentation and user groups. Because of their complexity, enterprise software ETL systems often have a higher price tag and require additional staff training and integration services to implement [35].

Open-Source ETL Tools: It is no wonder that open-source ETL solutions have entered the market with the development of the open-source movement. Many ETL solutions are now free and include graphical user interfaces for developing data-sharing procedures and monitoring information flow. The ability for enterprises to examine the tool's architecture and increase capabilities is a key advantage of open-source solutions. However, because they are not often sponsored by commercial businesses, open source ETL systems might differ in terms of upkeep, documentation, simplicity of use, and capability [36].

4.1 Some of the ETL Tools

This section discusses the latest tool for ETL. The overview of these tools is illustrated in Fig. 5 [37].

1. Hevo Data: Hevo Data is a platform that offers a simple solution for integrating and processing data from over 150 sources. It is designed to be easy to use, with a quick setup process that does not compromise on performance. With Hevo, users can import a wide range of data without the need for any coding skills. The platform provides strong connectivity to multiple sources, making it easy for users to bring in data in real-time [38]. Hevo has the following characteristics:

- Reliability at Scale: Hevo has a sturdy and scalable design that ensures data integrity and quick processing, even as the system expands.
- Monitoring and Observability: You can monitor the performance of your pipelines through easy-to-use dashboards that show all relevant statistics, and you can quickly get insight into your ELT processes using alerts and activity logs.
- StayinTotalControl:Hevoprovidesaflexiblesolutionthatadaptstoyourneeds,givingyoua variety of data ingestion options, control over ingestion frequency, JSON data parsing capabilities, a customizable destination workbench, and advanced schema management, giving you full control and versatility when automation is not enough.

- Transparent Pricing: Say farewell to complicated pricing plans and enjoy clarity with Hevo's straightforward pricing options. You can choose a package that suits your company's needs and stay in control of costs with spending alerts and pre-determined credit limits for sudden increases in data flow.

• Auto-Schema Management: Fixing incorrect schema after data has been imported into your warehouse can be challenging, but Hevo makes it easier by intelligently mapping the source schema to the target warehouse, avoiding schema errors.

24 × 7 Customer Support: Hevo offers more than just a platform, it provides a supportive partnership. You can find peace of mind with the platform's 24/7 live chat support and receive 24-h assistance during the 14-day free trial that features all the platform's capabilities.

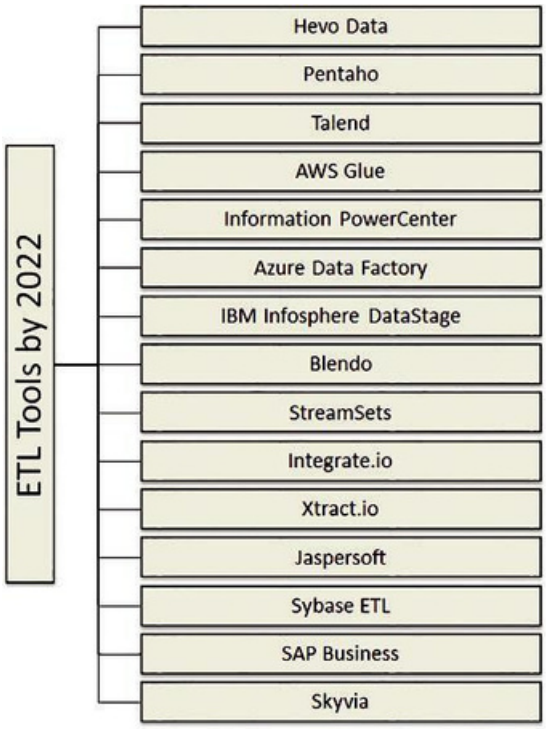

Figure 5: ETL tools by 2022

2. Pentaho: Pentaho is an important Business Intelligence tool that offers several key services and capabilities, such as OLAP, data integration, data mining, reporting, information dashboards, and ETL. Its purpose is to help users analyze complex data and turn it into valuable insights. Pentaho provides a wide range of report formats, including Text, HTML, CSV, XML, PDF, and Excel [39]. The following are the key features:

- Pentaho offers capabilities for processing and integrating data from various sources.
- Ithasafocusonmulti-cloudandhybridarchitecture
- Itisparticularlyusefulforon-premisebatchETLscenarios.
- One of its unique strengths is that it utilizes XML-based ETL methods, eliminating the need for coding and outperforming similar solutions.
- Itcanbedeployedeitheronacloudplatformoron-premise.

3. Talend: Talend provides comprehensive data management services, including data integrity, integration, governance, API creation, and integration with applications. It also works seamlessly with most cloud data warehousing solutions and is compatible with leading public cloud infrastructure providers [40]. The following are the main characteristics of Talend:

- Talend's strongest advantage is its ability to support hybrid and multi-cloud environments. This makes it a popular choice for customers who have very strict data protection requirements and need solutions that go beyond on-premise and cloud options.
- ThefunctionalityofTalendisgeneratedthroughacodegenerationprocess,meaningthatany changes in the logic require the code to be rewritten.
- Talendishighlycompatiblewithalargenumberofbothon-premiseandclouddatabases,aswell as various software-as-a-service solutions.
- The Talend Studio features a user-friendly graphical interface that enables the design of flow and transformation algorithms.
- Talendismosteffectivewithbatchprocedures.

4. AWS Glue: AWS Glue is a fully managed ETL service that simplifies the process of data preparation, ingestion, transformation, and catalog creation. It has all the necessary capabilities for data integration, so you can quickly start analyzing your data. With AWS Glue, data integration is made simple, and you can be up and running in minutes rather than months. AWS Glue provides both code-based and visual interfaces for easy data integration [41]. Additionally, the AWS Glue Data Catalog allows users to easily access and locate data. The main features of AWS Glue include:

- AWSGlueboastsseveralnoteworthyfeatures,includingautomaticdiscoveryofschemasandan integrated Data Catalog.
- Aserver-lessfull-fledgedETLPipelinemaybebuiltusingAWSGlueincombinationwith Lambda functions.
- AlthoughAWSGlueismainlydesignedforbatchprocessing,italsohasthecapabilitytosupport near real-time use cases through the use of Lambda functions.
- AWSGluefollowsapay-per-usepricingmodel,whereyouarechargedonanhourlybasiswith the billing done in second increments.

5. Information PowerCenter: Informatica PowerCenter is a robust and scalable enterprise data integration solution that covers all aspects of the data integration lifecycle. It offers data in batch, real-time, or change data capture modes, providing data on demand. As a unified platform, it can handle a wide range of data integration projects [21]. The key features of Informatica PowerCenter are:

- InformaticaPowerCenterisprimarilyabatch-orientedETLtoolwithconnectionstopopularcloud data warehouses such as DynamoDB and Amazon Redshift, among others.
- Itmeetssecurity,scalability,andcollaborationrequirementswithfeaturessuchasDynamic Partitioning, Metadata Management, Data Masking, and High Availability.
- InformaticaPowerCentermakesiteasiertocreateDWandDataMarts.

6. Azure Data Factory: Azure Data Factory is a server-less data integration solution that is completely managed. Without previous coding skills, you can easily design ETL methods in an easy environment with Azure Data Factory. The combined data may then be transferred to Azure Synapse Analytics to unearth critical insights that will help businesses succeed [42]. The following are the primary features of Azure Data Factory:

- AutonomousETLmaybeutilizedtoboostoperationalefficiencywhilealsoempoweringcitizen integrators.
- Azure Data Factory has more than 90 pre-built connectors, providing the ability to integrate all your software-as-a-service (SaaS) and software data.
- AzureDataFactoryhasbuilt-incontinuousintegration/continuousdelivery(CI/CD)andGit support, making it capable of swiftly migrating SQL Server Integration Services.
- AzureDataFactoryisaffordableduetoitspay-as-you-gopricingmodel.

7. IBM Infosphere DataStage: The IBM Infosphere DataStage ETL tool is a part of the IBM Infosphere and IBM Information Platform Solutions. It creates Data Integration solutions using graphical notation. IBM Infosphere DataStage is available in a variety of editions, including Enterprise, Server, and MVS [43]. It has the following characteristics:

- Containersandvirtualizationcanhelpyousavemoneyondatatransferwhilealsoallowingyouto enhance capabilities while protecting critical DataStage investments.
- IBM Infosphere DataStage is a batch-oriented ETL solution designed for large organizations with legacy data systems.
- Youcanrunanyjob30percentfasterwithaparallelengineandworkloadbalancing.
- WithIBMInfosphereDataStage,thedesignofETLjobscanbeeasilyseparatedfromtheir execution and transferred to any cloud environment.

8. Blendo: Blendo enables easy access to cloud data from various departments such as Marketing, Sales, Support, or Accounting, helping to boost data-driven business insights and growth. It features native data connections that simplify the ETL process, making it easier to automate data transforma-tion and management to speed up the process of gaining business intelligence insights [21]. The key features of Blendo are:

- Youcanconnecttoanydatasourceusingready-madeconnectors,savingyoucountlesshoursand allowing you to find significant insights for your company.
- YoucanlinkMailChimp,HubSpot,Salesforce,Stripe,Mixpanel,Shopify,GoogleAds,MySQL, and Facebook Ads, among many more, in minutes.
- Any SaaS application may be automated and synchronized into your Data Warehouse.
- Withreliabledata,analytics-readyschemas,andtablescreatedandoptimizedforanalysiswithany BI software, you can reduce the time it takes to get from discovery to insights.

9. StreamSets: You can leverage continuous data to power your digital transformation and modern analytics using the StreamSets DataOps platform. From a single login, you can monitor, develop, and execute intelligent Data Pipelines at scale. StreamSets allow for the quick creation and deployment of batch, streaming, machine learning, data centre, and ETL pipelines [21]. You may also manage and monitor all of your Data Pipelines through a single interface. The following are the major characteristics of StreamSets:

- Blendoallowsyoutoeliminategapsandblindspotsbyprovidingglobalvisibilityandcontrolover all data pipelines on a large scale across multi-cloud and hybrid environments.
- Blendo'shybridandmulti-clouddeploymentflexibilitymakesiteasytoswitchbetweenon-premises and multiple cloud environments without the need for additional work.
- YoucanuseStreamSetstokeeptasksgoingevenwhentheirstructuresandschemaschange.
- With automatic updates and no rewrites, maintenance time may be reduced by up to 80%.

10. Integrate.io: Integrate.io is recognized as a data integration and ETL tool that streamlines data processing, freeing up your organization to focus on insights instead of data preparation. It offers a user-friendly, coding-free interface through its point-and-click setup, making data integration and processing easier [44]. These are the key features of Integrate.io:

- Integrate.io is a user-friendly platform that can process millions of data points per minute with no lag.
- To provide a great user experience, Integrate.io provides unlimited video and phone support to all users.
- Integrate.io connects with over 140 various sources including databases, data warehouses, and cloud-based software as a service (SaaS) application. The platform provides data security measures, which can be paired with the Integrate.io Security Transformation capabilities to guarantee that your data remains secure and compliant.

11. Xtract.io: Xtract.io is a well-regarded tool for extracting web data that enables businesses to grow by utilizing AI-powered data aggregation and extraction. It offers a range of enterprise-grade platforms and solutions [45]. Xtract.io specializes in custom solutions that provide its clients with flexibility and adaptability. Additionally, Xtract.io provides precise location data, providing in-depth insights into your market, consumers, competitors, and products. These are the key features of Xtract.io.

- Xtract.iooffersadvanceddashboardsandreportsthatallowdecision-makersandanalyststo quickly make data-driven decisions with just a glance.
- Itcombinesdatafromseveralsources,removesduplicates,andimprovesit.Thismakesthe information easier to understand.
- Xtract.ioemploysAI/MLtechnologiessuchasImageRecognition,NLP,andPredictiveAnalytics to deliver precise information.
- Xtract.iodevelopsrobustAPIsthatofferasteadystreamofnewdatatoyourpremises.This encompasses both on-premises and cloud-based frameworks.

12. Jaspersoft: Jaspersoft is a highly respected name in the field of data integration, especially in regards to Extract, Transform, Load (ETL) processes. It is part of a larger business intelligence suite that provides a customizable and user-friendly platform to meet the specific needs of its customers [46]. The platform is known for its flexibility, adaptability, and ease of use for developers.

- Itisanopenarchitecturethatworksonanyplatform.Analyticsandreportswithprogrammatic control are easy to create, manage, and integrate.
- AllwebstandardsarefollowedbyJasper,includingitsJavaScriptAPIforembedding.Becauseof its API-first philosophy, it is a highly sought-after product in the business.
- Withmulti-tenantsupport,youcanmanagedatasecurityandresourceaccessforallofyourSaaS clients, and you may deploy in any way you choose.
- Itallowsyoutogeneratedatavisualizationsandreportsthatfollowstringentdesignguidelines.

13. Sybase ETL: Sybase ETL consists of two components: Sybase ETL Development and Sybase ETL Server. Sybase ETL Development is a graphical interface that allows for the creation and development of data transformation projects and tasks [47]. It features a simulation and debugging environment to streamline the development of ETL transformations. On the other hand, Sybase ETL Server is a scalable, distributed engine that connects to data sources and destinations, and extracts, transforms, and loads data using transformation processes.

- Sybase ETL enables the bulk transfer of data into a target database and supports the use of delete, update, and insert commands.
- SybaseETLhasthecapabilitytogatherdatafromavarietyofsources,includingSybaseIQ, Sybase ASE, Oracle, Microsoft Access, Microsoft SQL Server, among others.
- Sybase ETL allows for the cleansing, merging, transforming, and separating of data streams, which can then be utilized to perform insert, update, or delete operations on a target database.

14. SAP BusinessObjects Data Integrator: This data integration and ETL platform provides the ability to extract data from any source, convert and integrate it, and store it in any target database. Its primary function is to extract and modify data, and it includes basic tools for cleaning and organizing the data. Business rules and transformations can be established using a graphical user interface. This simplifies the execution of your operations [48]. The following are the key features:

- ItiscompatiblewiththeplatformsWindows,SunSolaris,Linux,andAIX.
- ItmaybeusedtodevelopanytypeofDWorDataMart.
- BatchjobsmaybeexecutedscheduledandmonitoredusingSAPBusinessObjectsDataIntegrator.

15. Sky Via: Sky via is a cloud platform that offers cloud-to-cloud backup, Data interface data access, SQL administration and data integration without scripting. Sky via is incredibly scalable since it offers configurable pricing choices for each product, making it suitable for all types of enterprises ranging from large corporations to small startups [38]. It also offers current Cloud agility by removing the need for manual upgrades or deployment. The following are Skyvia's key characteristics:

- YoucaneasilyautomatedatagatheringfrommanycloudsourcestoaDataWarehouseordatabase, and you can migrate your company's data between cloud apps with a few clicks.
- SkyviaalsoincludestemplatesforcommonlyencounteredDataIntegrationscenarios.
- Skyviaenablesyoutomaintainsourcedatarelationshipsinthedestination,aswellasdataimport without duplication and bi-directional synchronization.

4.2 ETL Tools Evaluation

To assess ETL technologies, consider the following factors: the complexity of your ETL require- ments, current cloud vendor partnerships, and in-house development skills. When determining the complexity of your ETL requirements, consider cloud service connections, structured and unstruc-tured data, and interaction with the source data platform [49]. You must choose ETL solutions based on your requirements. For example, a large corporation would have different data processing requirements than a small software startup. If you are contemplating cloud tools, you should also examine current cloud vendor connections. If your team can efficiently manage and use any open source ETL tools, then there are further solutions that may be highly cost-effective and suit your ETL demands for enterprises with considerable in-house development skills [50].

Furthermore, each organization has its own business strategy and culture, which will be reflected in the data that a firm gathers and appreciates. However, there are common criteria against which ETL technologies may be measured that will be applicable to any organization, as illustrated in Fig. 6.

Use Case: A crucial factor for ETL solutions is the use case. If your firm is tiny or your data analysis needs are simple, you may not require a solution as robust as large enterprises with complicated datasets.

Budget: Another key consideration while selecting ETL software is the cost. Although open-source technologies are often free to use, they may lack the functionality and support of enterprise-grade

products. Another factor to consider is the resources required to hire and retain developers if the software requires a lot of coding.

Capabilities: The finest ETL solutions may be tailored to match the data requirements of various teams and business processes. One method ETL systems may ensure data quality and decrease the effort necessary to examine datasets is through automated features such as de-duplication. Furthermore, data connectors simplify platform sharing.

Data Sources: The ideal ETL solution should be able to access data from any location, whether it be on-premise or in the cloud, and handle diverse data structures, including both structured and unstructured data. It should also be able to collect data from various sources and convert it into a standardized format for storage.

Technical Literacy: The proficiency of developers and end users in handling data and code is crucial. Ideally, the development team should have expertise in the programming languages used by the tool, especially if manual coding is required. On the other hand, if users are not skilled in writing complex queries, a tool that automates this process would be a great fit.

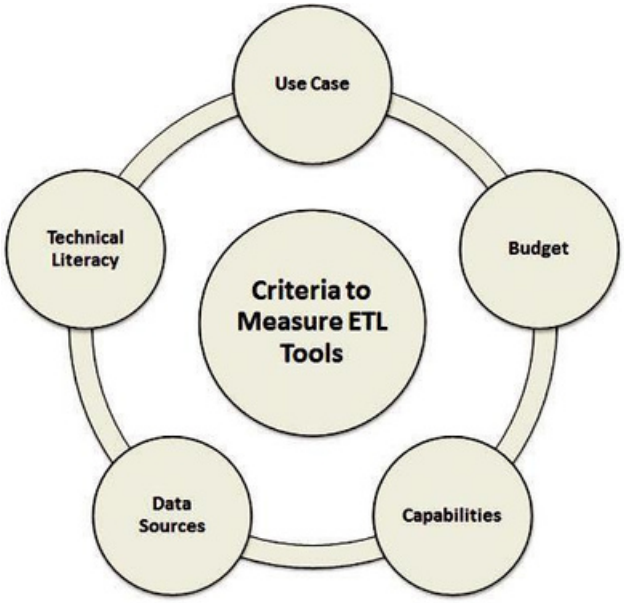

Figure 6: Common criteria to measure ETL tools

5 Conclusion

There are various processes involved in DW, and ETL is the main process in DW. It is designed to save cost and time when a new DW or data mart is developed. There is various tool in techniques used in ETL processes and testing own with its strengths and weaknesses. This study presents the overview of different state-of-the-art ETL tools, techniques, testing and evaluation criteria. This study can be the base line for new users in this area. The conclusion underscores the significance of ETL in the successful implementation of data warehousing projects. It is essential for professionals and researchers to have a thorough understanding of ETL, including its components, techniques, tools, and evaluation criteria. This study serves as a valuable resource in this regard, providing comprehensive information on ETL that can be used to guide the development and implementation

of effective data warehousing projects. This study also highlights the importance of the ETL process in data warehousing and provides a comprehensive overview of the various components of ETL, its techniques, tools, and evaluation criteria. This information is essential for professionals and researchers looking to effectively implement data warehousing projects and ensure the quality and reliability of the data in the data warehouse. This research study provides a thorough exploration of ETL methods, tools, procedures, and assessments in data warehousing. It covers everything from the basic function of ETL in data integration to the complexities of ETL testing, tool kinds, and evaluation standards. It highlights how important it is for ETL to effectively integrate data from many sources in order to provide business intelligence and decision-making.

The future research will explore advanced ETL automation through ML and AI-driven processes; it will also look into real-time ETL solutions to address dynamic data integration needs; it will delve deeper into the security and privacy considerations surrounding ETL; it will address scalability issues in the context of growing data volumes; it will promote industry standards and best practices for ETL; and it will conduct in-depth case studies and empirical validations of the ETL techniques and tools covered in the paper. Through increased efficiency, security, and flexibility in the quickly changing data landscape, these research directions will support the ongoing development and adaption of ETL operations in data warehousing.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Saifullah Jan, Bilal Khan; data collection: Bilal Khan; analysis and interpretation of results: Saifullah Jan, Wahab Khan; draft manuscript preparation: Wahab Khan, Muhammad Imran Chughtai. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.