International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

EXTENSIVE STUDY OF ETL and ETL-Tools

Mangalam Palod*, Deepika Dash**

*(Department of Computer Science and Engineering, RV College of Engineering, Bangalore
Email: mangalam.palod@gmail.com)

**(Department of Computer Science and Engineering, RV College of Engineering, Bangalore
deepikadash@rvce.edu.in)

----------**********----------

Abstract:

ETL is a process that extracts data from a variety of sources, transforms it (using computations, concatenations, and other operations), and puts it into a Data Warehouse system.
ETL tools are a type of software that extracts data from a variety of sources, cleans, modifies, transforms, and merges it before storing it in a data warehouse.
A data warehouse collects data from multiple operational or external systems in order to deliver integrated and understandable information to its end-users.
Building of a data warehouse requires the execution of the Extraction-Transformation-Load (ETL) process and has a huge market impact.
The ETL process in data warehousing is in charge of extracting data from operational systems and storing it in the data warehouse. Constructing the ETL process is one of the most difficult aspects of warehouse construction. We will attempt to discuss the ETL process, and ETL tools methodologies in this document. This paper studies ETL-Tools to perform ETL and gives an insight on different platforms.

Keywords — **ETL, ELT, ETL Tools, Dataware house.**

----------**********----------

I. INTRODUCTION

To fulfil the aim of aiding business analysis, one must routinely load the data warehouse. To use it, data must be retrieved and transferred into the warehouse from one or more operating systems. ETL describes the process of extracting data from source systems and moving it into the data warehouse. Because it omits the transportation step and suggests that each of the other parts of the process is different, the acronym ETL may be oversimplified. ETL refers to the complete process. This study examines the major ETL technologies, such as data extraction, data transformation, incremental data loading, and break-points transmission.

Data warehouse is subject Oriented, Integrated, Time-Variant and non-volatile collection of data that supports decision making process in an organization[1]. The operational database is subjected to a variety of daily transactions, making data analysis more difficult and time intensive. Data must be exchanged throughout applications or systems in order to integrate them and provide at least two apps with the same view of the world. The majority of this data exchange was handled by processes akin to what we now refer to as ETL. One of the most complex and resource-intensive aspects of a data warehouse project is building and maintaining the ETL process. ETL tools are used to handle this process in many data warehousing initiatives. For example, AWS glue offers ETL

capabilities while also utilising database capabilities. Some construct their own ETL tools and techniques, either into or out of the database.

Aside from extraction, transformation, and loading, there are a few other tasks that are critical for a successful ETL deployment as part of the data warehouse's everyday operations and future upgrades. Data is extracted from multiple, heterogeneous data sources, transformed, and then loaded into the data warehouse using the ETL tool.

II. ETL

ETL consists of three processes that allow data to be sent from source to destination.

A. Extraction

The system determines the required data sources as well as the fields to be extracted from the sources, during the extract phase of ETL. Before data can be moved to a new site, it must first be retrieved from its original place.

In this stage of the ETL system, data is taken from the source system and placed in the staging area.Data may be obtained in its raw form from a number of sources, including older systems and databases.

B. Transformation

Rules and regulations that assure data quality and accessibility can be introduced at this stage of the ETL process, as well as a set of functions that can be applied to the data that has been extracted.

The data transformation process is divided into several sub-processes: Cleaning entails correcting data inconsistencies and missing values. The term "standardisation" refers to the process of applying rules to a data set. The process of deleting or eliminating redundant data is known as deduplication. During verification, unusable data is removed, and abnormalities are identified. Sorting is the process of categorising data.

The most significant phase in the ETL process is transformation, which is usually regarded as the most important. Data transformation improves data integrity and guarantees that data arrives at its new place fully compliant and ready to use.

In its current condition, the data collected from the source server is useless.

C. Loading

As the final stage in the ETL process, the converted data is imported into a new destination. Data can be loaded in bulk (full load) or at predefined intervals (interval loading) (incremental load).

In an ETL full loading scenario, everything that comes off the transformation assembly line is turned into new, unique entries in the data warehouse. Though this is sometimes useful for study, comprehensive loading leads in data sets that grow rapidly and become difficult to manage.

A big volume of data must be put into a data warehouse in a short amount of time (nights). As a result, the loading procedure should be sped up.

Incremental loading is a way of loading that is less complete but more customizable. Incremental loading compares fresh data to what's currently on hand, and only produces new entries if it finds fresh and unique data.

III. ETL V/S ELT

It is Extract, Load and Transform. The steps are same as ETL, just the order of execution is different in ELT. The difference between ETL and ELT is also important for understanding ETL Tools.

TABLE I DIFFERENCE BETWEEN THE ETL VS ELT

	ETL	ELT
System Flow	Within a staging region, data transformations take place immediately after extraction. The data is placed into the data warehouse after it has been transformed.	The data is extracted first, and then fed into the target data system. Only afterwards, for analytical purposes, is part of the data changed on a "as-needed" basis.
Processing Location	Before transmitting the data to the Datawarehouse DB, the staging server changes it.	Data is always stationed at database.
Datasets	Smaller data sets that require complicated processing are best served by ETL.	When dealing with large amounts of structured and unstructured data, ELT is the best option.
Processing style and Analysis	Views are created using numerous scripts, therefore deleting a view involves destroying data.	Creating and maintaining ad hoc views is inexpensive.
Maintenance	High maintenance-must choose which data to load and transform, and must do it again if destroyed or if the main data repository is to be enhanced.	No maintenance, since complete data is available.

IV. ETL TOOLS

A. XPLENTY

Xplenty is a cloud-based ETL and ELT data integration tool that allows you to connect many data sources with ease. This platform provides a simple and straightforward visual interface for creating data pipelines between multiple sources and destinations. It includes more than 100 major data stores and SaaS apps. MySQL, MongoDB, PostgreSQL, Amazon Redshift, Google Cloud Platform, Facebook, QuickBooks, and dozens of other applications are on the list[4]. It's other benefits include scalability, security, and outstanding customer service. Field Level Encryption, for example, is a new feature in Users may encrypt and decode data fields using their own encryption key using the tool. It also ensures regulatory compliance with legislation such as HIPPA, GDPR, and the CCPA.

B. Talend

Talend Data Integration is, an open-source ETL (enterprise data warehousing) data integration tool, a free open-source solution that simplifies ETL testing[6]. It comes with all ETL testing features as well as a continuous delivery method. The user can run ETL jobs on a remote server with a range of operating systems using Talend Data Integration Tool. ETL testing guarantees that data is translated without loss from the source system to the target system and that the transformation rules are followed. The Talend platform works with both on-premises and cloud data sources and comes with hundreds of pre-built integrations. Any relational database, flat files, and other file types are supported by Talend Data Integration. The ETL process is simplified and developed with the help of an integrated GUI. We can identify flaws at an initial stages with the help of Talend, which helps save money. Talend is capable of swiftly detecting business uncertainty and consistency in transformation rules. Switching is possible in Talend. With thorough performance statistics, Talend can track real-time flow of data.

C. Informatica PowerCenter Tool

Informatica PowerCenter is an enterprise data integration platform for ETL jobs. Informatica's PowerCenter is simply one of several cloud data management technologies in the Informatica suite. PowerCenter has a reputation for great performance and compliance with a wide range of data sources, including SQL and non-SQL databases, as an enterprise-class, database-neutral solution[3]. It's a batch-based tool. Informatica PowerCenter has some drawbacks, including high cost and a hard learning curve, which may prevent smaller firms with limited technical expertise. Informatica uses an ETL architecture to process data. Workflow Manager, Monitor, Designer, and Repository Manager are the four main components of PowerCenter. A data engineer must employ each of these components in a complex yet systematic design sequence to create a full ETL pipeline. It also has a cloud counterpart that allows users to access repositories on the business premises and perform transformation operations in the cloud. It's change connectors support AWS DynamoDB, AWS Redshift, and other popular cloud data warehouses. It also supports a number of data storage and software-as-a-service options. Because of obligatory compliance requirements, Informatica PowerCenter is better suited for enterprises that require enterprise-grade security and data governance within their on-premise data. Even Informatica PowerCenter's cloud technology is better suited for on-premise data, with a focus on data protection[2].

D. Pentaho

Pentaho is designed for on-premise, batch ETL use cases. It provides data integration and processing capabilities from a variety of data sources. Pentaho also places a strong bet on hybrid cloud and multi-cloud architectures. Pentaho relies on the interpretation of ETL methods stored in XML format to function. Pentaho is superior than Talend for ad-hoc analysis because it does not require code generation. Pentaho does not provide cost information up front. Developers can't always figure out what's causing the error because there aren't any specific explanations on the logging screen. When enterprises opt for open source ETL technologies in an on-premise ecosystem, Pentaho is frequently used.

E. AWS Glue

AWS offers Glue, a cloud-based real-time ETL solution that is available on a pay-as-you-go basis. AWS glue is largely batch-oriented, although it can also enable lambda-based near-real-time use cases. If the majority of the data sources from which you want to ingest data are on AWS, Glue makes it simple to ETL the data. Outside of the AWS ecosystem, support for sources and destinations is limited[5]. Glue provides various unique features, such as an integrated data catalogue and automated schema discovery. It can construct a serverless full-fledged ETL pipeline using AWS Glue and lambda functions. Glue also offers Dev Endpoints and Notebooks, which provide a dedicated Spark cluster on which you can run jobs indefinitely, making it easier to build and test scripts. Because I am unable to use them in my development environment, I am unable to comment on their functionality.

F. QuerySurge

The QuerySurge tool was created to test Data Warehouses and Big Data. It also ensures that the data collected and loaded from the source system to the destination system is accurate and formatted correctly. With QuerySurge, any errors or differences can be instantly identified. QuerySurge is an ETL and Big Data testing tool that is automated. It improves data quality and speeds up testing cycles. The Query Wizard is used to validate data. It saves time and expense by automating the manual efforts and schedule test for a specified period. ETL testing is supported by QuerySurge for a variety of platforms, including IBM, Oracle, Microsoft, and others. It enables the creation of test scenarios and test suits, as well as customised reports, without the need for SQL knowledge. It creates the email using an automated method. Via the ETL process, QuerySurge verifies, converts,

Table II . Result

	XPlenty	Talend	Informatica PowerCenter Tool	Pentaho	AWS Glue	QuerySurge
ETL	✔	✔	✔	✔	✔	✔
ELT	✔	✔	✔	✔	×	×
Real-Time	×	✔	✔	×	✔	×

and upgrades data. It's a commercial solution that uses the ETL procedure to link sources and enhance data.

V. RESULTS

The results are depicted in Table 2 above.

VI. CONCLUSIONS

We explored both the ETL and ELT approaches for loading data into a data warehouse in this study. This study discusses the advantages and disadvantages of both techniques.The paper also showed which technique is better than the others and in which conditions.

All tools have their own set of strengths and disadvantages, some of them has the advantage due to its maturity and reliability, as well as its suitability for enterprise-scale deployments, excellent support, speed, implementation, and ease of usage. Despite the fact that some of them are free, its lack of support, documentation, and large-scale deployments making it unsuitable for commercial use, particularly with financial and cloud-based systems.

ACKNOWLEDGMENT

I would like to acknowledge the support provided by teacher's of Department of Computer Science and Engineering, RV College of Engineering, Bangalore, India through their assistance in the research work.

REFERENCES

[1] Barateiro, J., & Galhardas, H. (2005). *A Survey of Data Quality Tools. Datenbank-Spektrum* 14, 15-21
[2] Kimbal, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses.* John Wiley & Sons.
[3] Informatica PowerCenter. Product's web page at *http://www.informatica.com/products_services/powercenter/Pages/index.aspx*
[4] Xplenty. Product's web page at *https://www.xplenty.com/blog/top-7-etl-tools*
[5] AWSGLUE. Products Web Page at *https://aws.amazon.com/glue/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc*
[6] Talend, Products web page at *https://www.talend.com/*
[7] Carreira P., Galhardas, H., Pereira, J., Martins, F., & Silva, M. (2007*). On the performance of one-to-many data transformations. Proceedings* of the Fifth International Workshop on Quality in Databases (QDB 2007), pp.: 39-48, in conjunction with the VLDB 2007 conference, Vienna, Austria, September 23, 2007

