CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

APEX INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Introduction to Data Science (21CST-292)

Faculty: Dr. Jitender Kaushal (E14621)

Associate Professor

Lecture - **ETL and ELT**

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

The Course aims to:

- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

COURSE OUTCOMES

On completion of this course, the students shall be able to:-

CO3	Illustrate and compare various types of analytics techniques.

ETL vs ELT Overview

The ETL and ELT acronyms both describe processes of cleaning, enriching, and transforming data from a variety of sources before integrating it for use in data analytics, business intelligence and data science.

ETL vs ELT Comparison

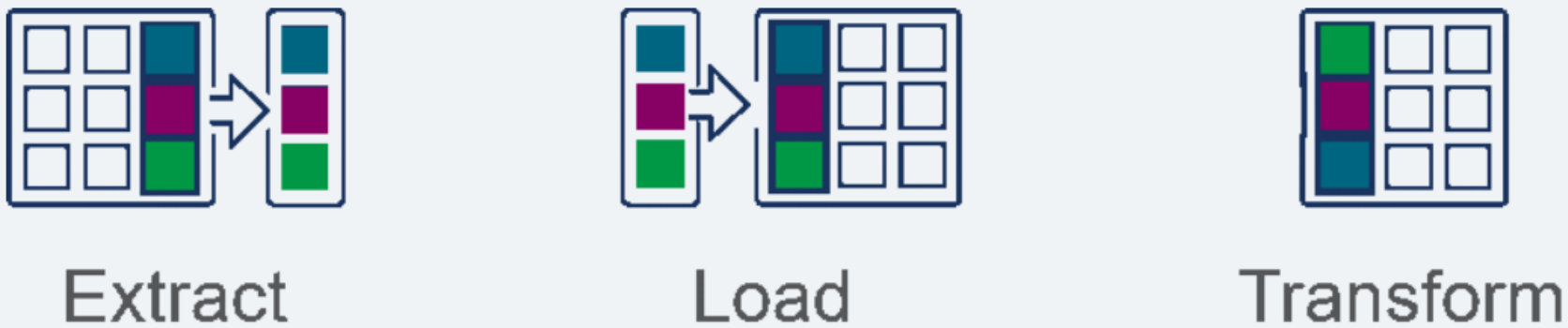
- **The letters stand for Extract, Transform, and Load.**
- **Extract** refers to the process of pulling data from a source such as an SQL or NoSQL database, an XML file or a cloud platform.
- **Transform** refers to the process of converting the format or structure of a data set to match that of a target system.
- **Load** refers to the process of placing a data set into a target system.

ETL stands for Extract > Transform > Load

In the ETL process, data transformation is performed in a staging area outside of the data warehouse and the entire data must be transformed before loading. As a result, transforming larger data sets can take a long time up front but analysis can take place immediately once the ETL process is complete.

ELT stands for Extract > Load > Transform

In the ELT process, data transformation is performed on an as-needed basis in the target system itself. As a result, the transformation step takes little time but can slow down the querying and analysis processes if there is not sufficient processing power in the cloud solution.



There is a place for each process.

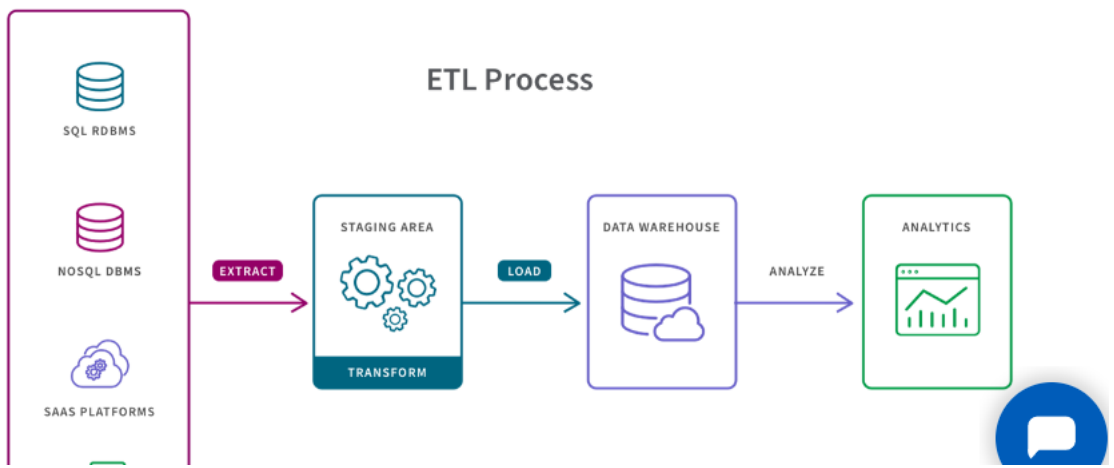
The ETL process is appropriate for small data sets which require complex transformations. The ELT process is more appropriate for larger, structured and unstructured data sets and when timeliness is important.

What is ETL?

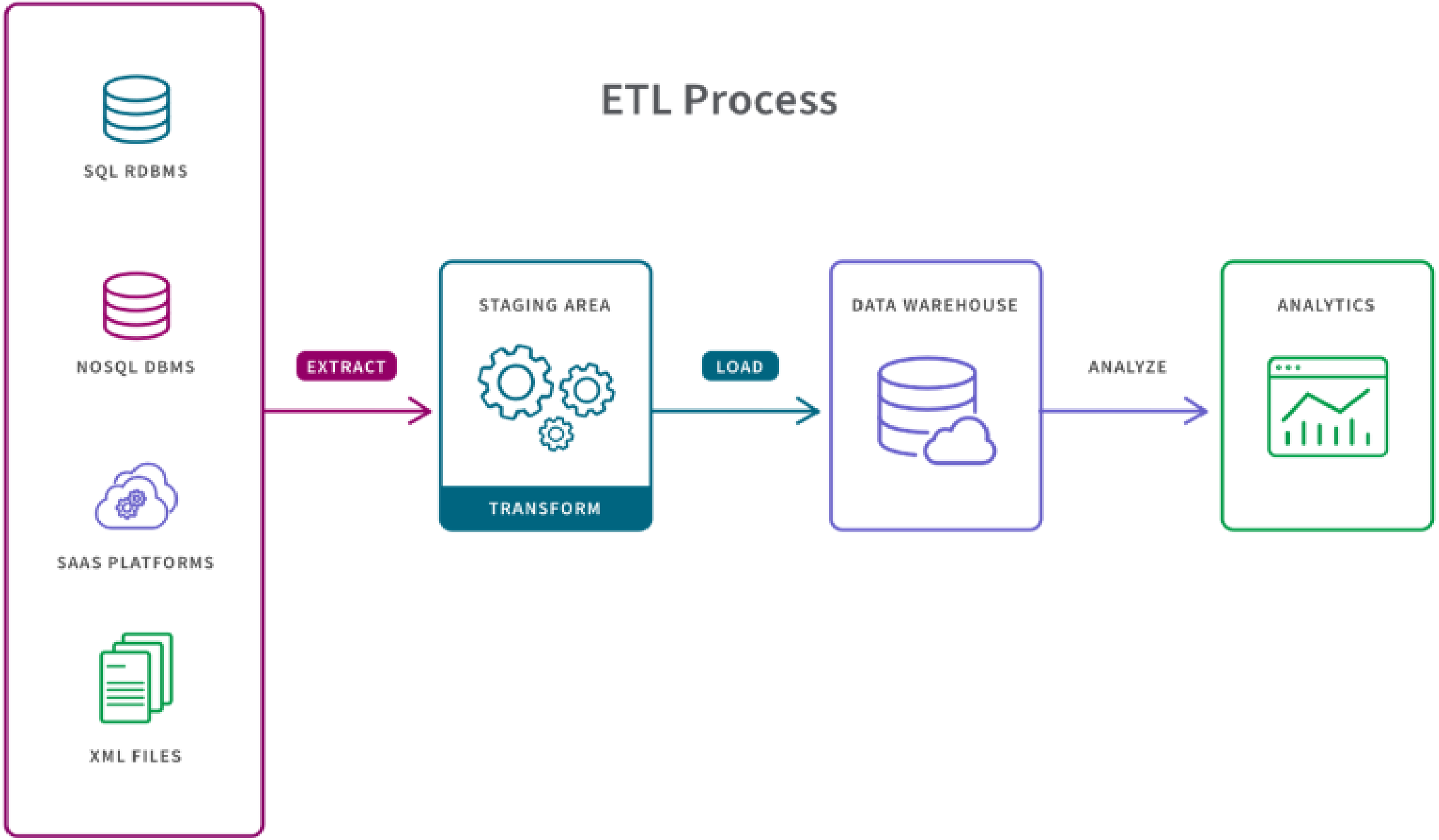
ETL is an acronym for “Extract, Transform, and Load” and describes the three stages of the traditional data pipeline. The ETL process is appropriate for small data sets which require complex transformations.

ETL Process

1. A predetermined subset of data is extracted from the source.
2. Data is transformed in a staging area in some way such as data mapping, applying concatenations or calculations. Transforming the data before it is loaded is necessary to deal with the constraints of traditional data warehouses.
3. Data is loaded into the target data warehouse system and is ready to be analyzed by BI tools or data analytics tools.



ETL Process


SQL RDBMS
NOSQL DBMS
SAAS PLATFORMS
XML FILES
EXTRACT
STAGING AREA
TRANSFORM
LOAD
DATA WAREHOUSE
ANALYZE
ANALYTICS

Key benefits of ETL

- **Data analysis** on a single, pre-defined use case can be slightly more stable and faster with the ETL process given that the data set has already been structured and transformed.
- **Compliance** with GDPR, HIPAA, and CCPA standards is easier with ETL given that users can omit any sensitive data prior to loading in the target system.

The **General Data Protection Regulation Act of 2016 ('EU GDPR') and the California Consumer Privacy Act of 2018 ('CCPA')** both aim to give strong protection for individuals regarding their personal data collected for business use, or share consumer data, whether the information is obtained online or offline.

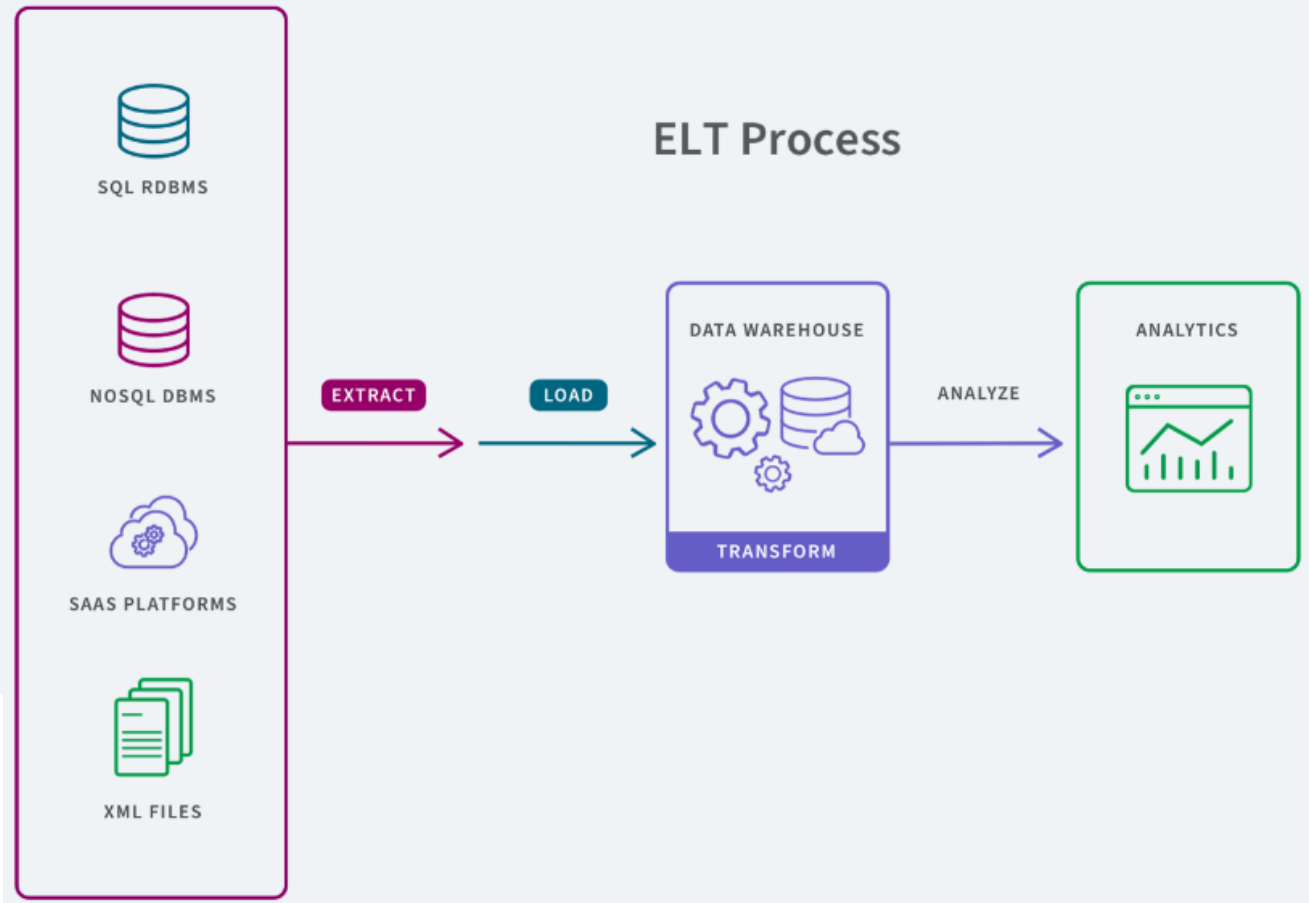
The **Health Insurance Portability and Accountability Act** (HIPAA) was enacted in 1996 by the US government and provides the rules and regulations for protecting privacy of Patient Health Information.

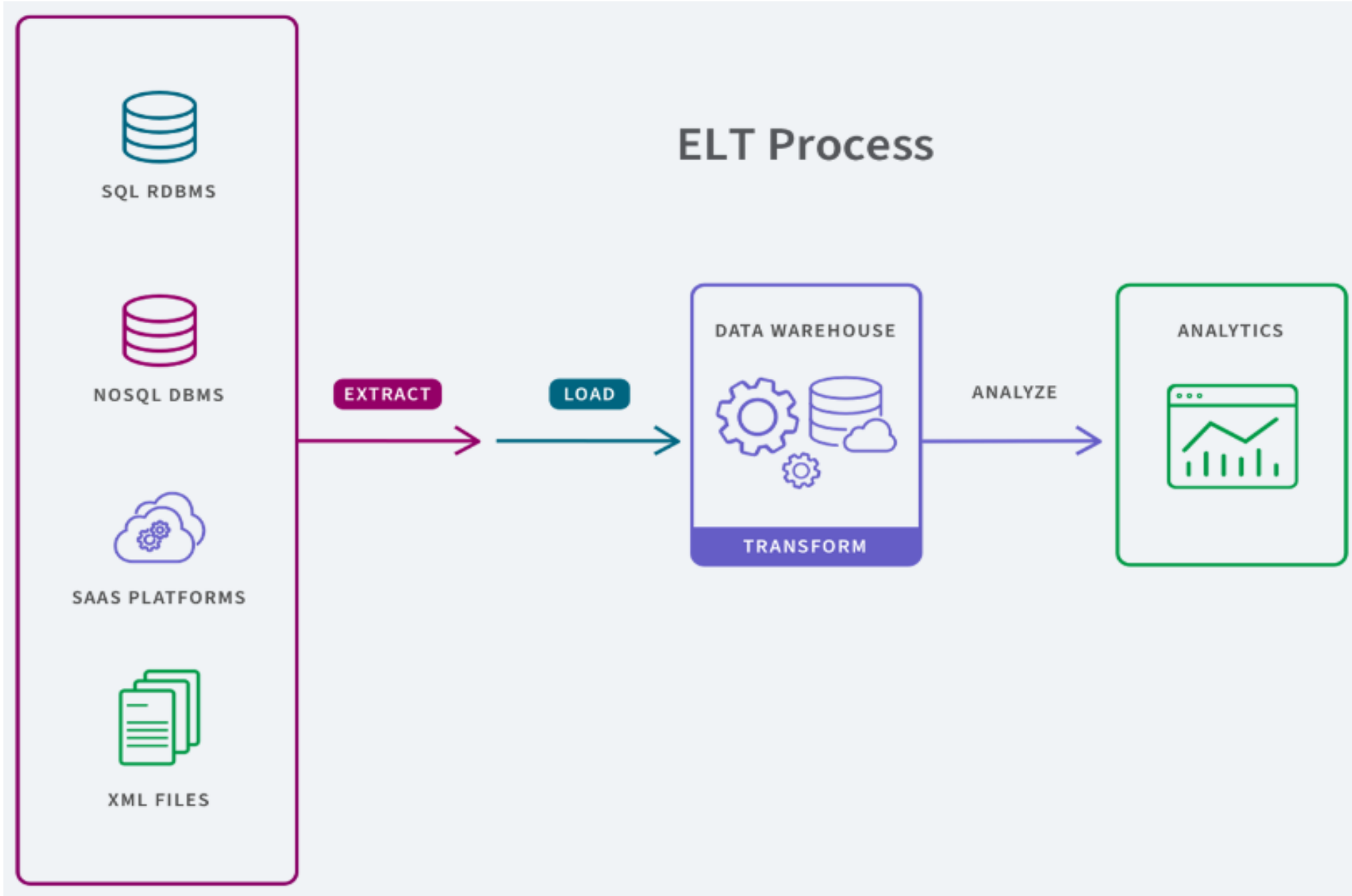
What is ELT?

ELT is an acronym for “Extract, Load, and Transform” and describes the three stages of the modern data pipeline. The ELT process is more cost effective then ETL, is appropriate for larger, structured and unstructured data sets and when timeliness is important.

ELT Process

1. All data is extracted from the source.
2. All data is immediately loaded into the target system (either a data warehouse, data mart, or data lake). This can include raw, unstructured, semi-structured and structured data types.
3. Data is transformed in the target system and is ready to be analyzed by BI tools or data analytics tools




ELT Process
SQL RDBMS
NOSQL DBMS
SAAS PLATFORMS
XML FILES
EXTRACT
LOAD
DATA WAREHOUSE
TRANSFORM
ANALYZE
ANALYTICS

Key benefits of ELT

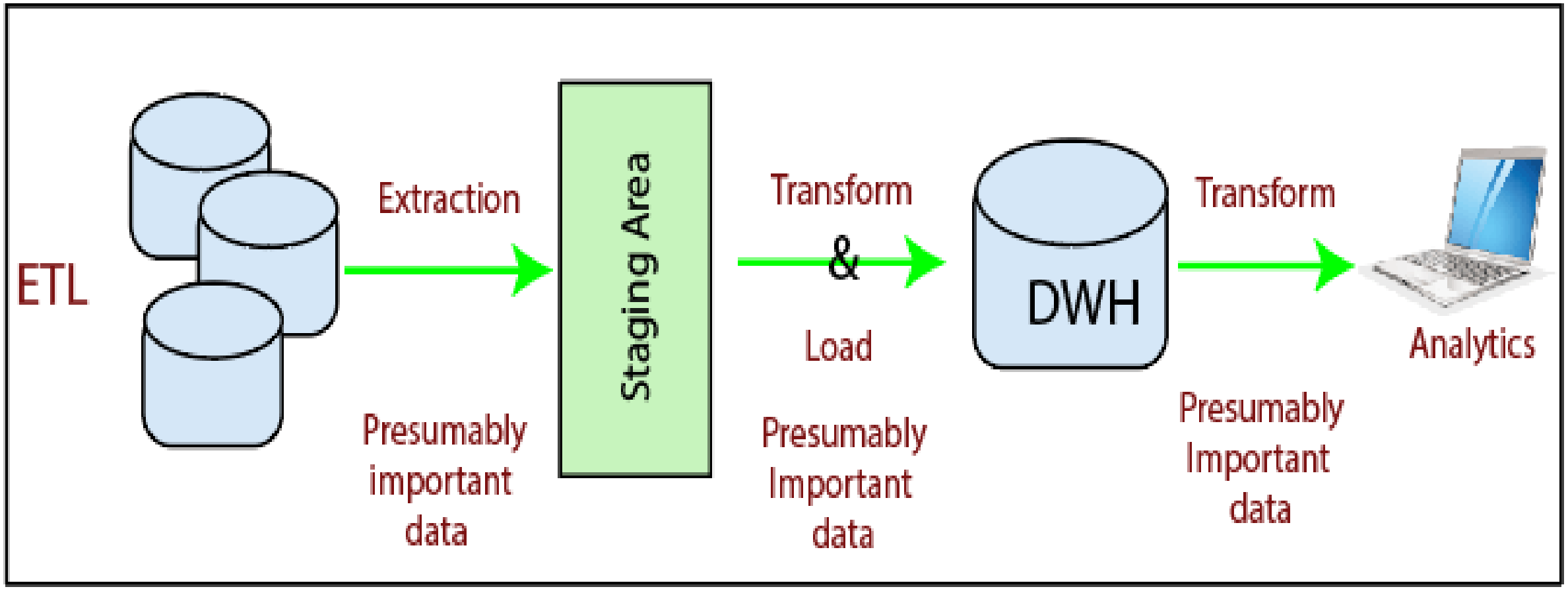
- **Real-time, flexible data analysis.** Users have the flexibility to explore the complete data set, including real-time data, in any direction, without having to wait for IT to extract, transform and load more data.
- **Lower cost and lower maintenance.** ELT benefits from a robust ecosystem of cloud-based platforms which offer much lower costs and a variety of plan options to store and process data. And, the ELT process typically requires low maintenance given that all data is always available and the transformation process is usually automated and cloud-based.

OR

YOU MAY REFER TO THE FOLLOWING CONTENT

ETL (Extract, Transform, and Load)

Extract, Transform and Load is the technique of **extracting the record** from **sources** (which is present outside or on-premises, etc.) **to** a **staging area, then** *transforming or reformatting* with **business manipulation** performed on it, in order **to fit** the **operational needs** or **data analysis**, and **later loading into** the goal or **destination** databases or data **warehouse**.


ETL
Extraction
Presumably
important
data
Staging Area
Transform
&
Load
Presumably
Important
data
DWH
Transform
Presumably
Important
data
Analytics

Strengths

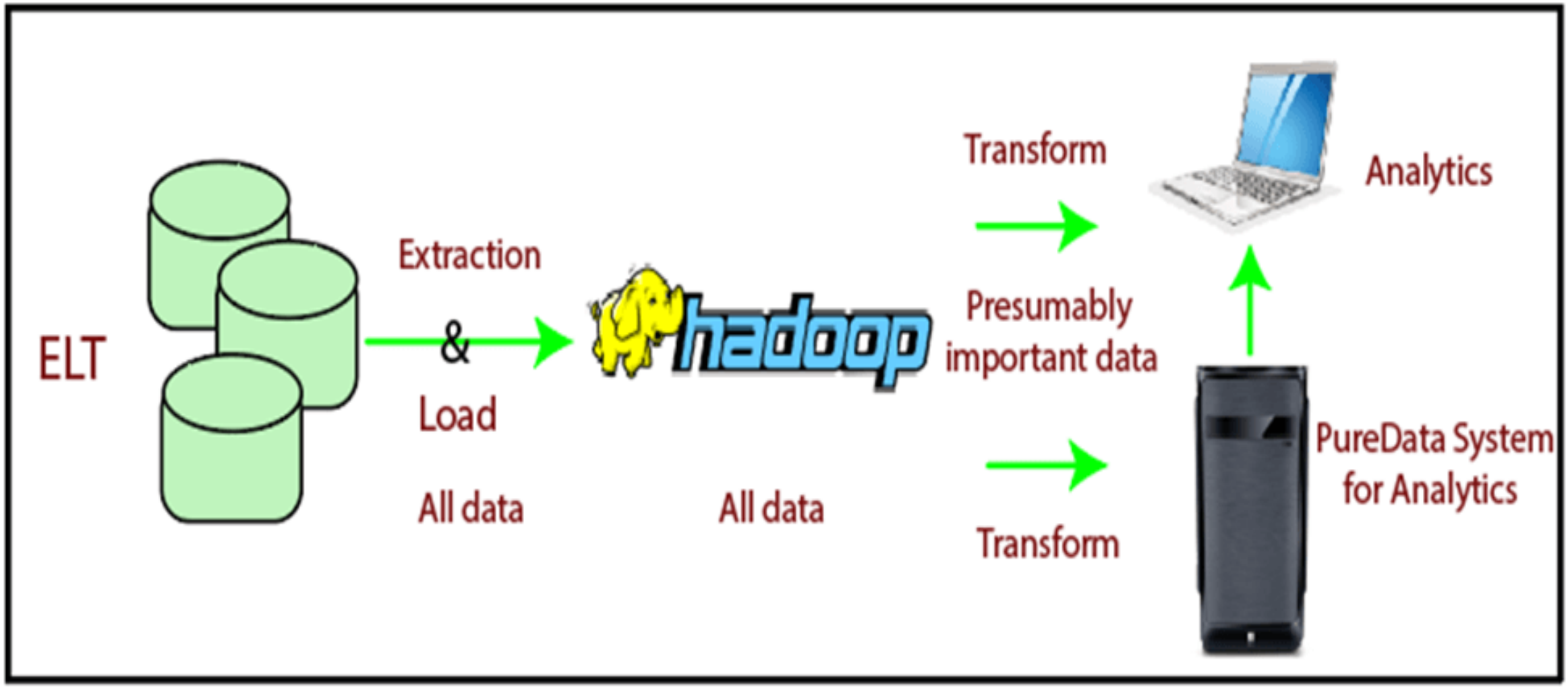
- **Development Time:** Designing from the output backwards, provide that only information applicable to the solution is extracted and processed which is potentially decreasing development, and processing overhead.
- **Targeted data:** Due to the targeted feature of the load process, the warehouse contains only information relevant to the presentation. Reduced warehouse content simplify the security regime enforce and hence the administration overheads.
- **Tools Availability:** The number of tools available that implement ETL provides the flexibility of approach and the opportunity to identify the most appropriate tool. The proliferation of tools has to lead to a competitive functionality war, which often results in loss of maintainability.

Weaknesses

- **Flexibility:** Targeting only relevant information for output means that any future requirements that may need data that was not included in the original design will need to be added to the ETL routines. Due to the nature of tight dependency between the methods developed, this often leads to a need for fundamental redesign and development. As a result, this increase the time and cost involved.
- **Hardware:** Most third-party tools utilize their engine to implement the ETL phase. Regardless of the estimate of the solution, this can necessitate the investment in additional hardware to implement the tool's ETL engine. The use of third-party tools to achieve the ETL process compels the information of new scripting languages and processes.
- **Learning Curve:** Implementing a third-party tools that uses foreign processes and languages results in the learning curve that is implicit in all technologies new to an organization and can often lead to consecutive blind alleys in their use due to shortage of experience.

ELT (Extract, Load and Transform)

ELT stands for Extract, Load and Transform is the various sight while looking at data migration or movement. ELT involves the extraction of aggregate information from the source system and loading to the target method instead of transformation between the extraction and loading phase. Once the data is copied or loaded into the target method, then change takes place.


ELT
Extraction
&
Load
All data
hadoop
All data
Transform
Presumably
important data
Transform
Analytics
PureData System
for Analytics

The **extract** and **load** step can be isolated from the transformation process. Isolating the load phase from the transformation process delete an inherent dependency between these phases. In addition to containing the data necessary for the transformations, the extract and load process can include components of data that may be essential in the future. The load phase could take the entire source and loaded it into the warehouses.

Separating the phases enables the project to be damaged down into smaller chunks, thus making it more specific and manageable.

Performing the data integrity analysis in the staging method enables a further phase in the process to be isolated and dealt with at the most appropriate point in the process. This method also helps to ensure that only cleaned and checked information is loaded into the warehouse for transformation.

Isolating the transformations from the load steps helps to encourage a more staged way to the warehouse design and implementation.

Strengths

- **Project Management:** Being able to divide the warehouse method into specific and isolated functions, enables a project to be designed on a smaller function basis, therefore the project can be broken down into feasible chunks.
- **Flexible & Future Proof:** In general, in an ELT implementation, all record from the sources are loaded into the data warehouse as part of the extract and loading process. This, linked with the isolation of the transformation phase, means that future requirements can easily be incorporated into the data warehouse architecture.
- **Risk minimization:** Deleting the close interdependencies between each technique of the warehouse build system enables the development method to be isolated, and the individual process design can thus also be separated. This provides a good platform for change, maintenance and management.
- **Utilize Existing Hardware:** In implementing ELT as a warehouse build process, the essential tools provided with the database engine can be used.
- **Utilize Existing Skill sets:** By using the functionality support by the database engine, the existing investment in database functions are re-used to develop the warehouse. No new skills need to be learned, and the full weight of the experience in developing the engine? technology is utilized, further reducing the cost and risk in the development process.

Weaknesses

- **Against the Norm:** ELT is a new method to data warehouse design and development. While it has proven itself many times over through its abundant use in implementations throughout the world, it does require a change in mentality and design approach against traditional methods.
- **Tools Availability:** Being an emergent technology approach, ELT suffers from the limited availability of tools.

Difference between ETL vs. ELT

Basics	ETL	ELT
Process	Data is transferred to the ETL server and moved back to DB. High network bandwidth required.	Data remains in the DB except for cross Database loads (e.g. source to object).
Transformation	Transformations are performed in ETL Server.	Transformations are performed (in the source or) in the target.
Code Usage	Typically used for Source to target transfer Compute-intensive Transformations Small amount of data	Typically used for High amounts of data
Time-Maintenance	It needs highs maintenance as you need to select data to load and transform.	Low maintenance as data is always available.
Calculations	Overwrites existing column or Need to append the dataset and push to the target platform.	Easily add the calculated column to the existing table.

Parameters	ETL	ELT
Process	Data is transformed at staging server and then transferred to Datawarehouse DB.	Data remains in the DB of the Datawarehouse.
Code Usage	Used for •Compute-intensive Transformations •Small amount of data	Used for High amounts of data
Transformation	Transformations are done in ETL server/staging area.	Transformations are performed in the target system
Time-Load	Data first loaded into staging and later loaded into target system. Time intensive.	Data loaded into target system only once. Faster.
Time-Transformation	ETL process needs to wait for transformation to complete. As data size grows, transformation time increases.	In ELT process, speed is never dependent on the size of the data.

Parameters	ETL	ELT
Time- Maintenance	It needs highs maintenance as you need to select data to load and transform.	Low maintenance as data is always available.
Implementation Complexity	At an early stage, easier to implement.	To implement ELT process organization should have deep knowledge of tools and expert skills.
Support for Data warehouse	ETL model used for on-premises, relational and structured data.	Used in scalable cloud infrastructure which supports structured, unstructured data sources.
Data Lake Support	Does not support.	Allows use of Data lake with unstructured data.
Complexity	The ETL process loads only the important data, as identified at design time.	This process involves development from the output-backward and loading only relevant data.
Cost	High costs for small and medium businesses.	Low entry costs using online Software as a Service Platforms.
Lookups	In the ETL process, both facts and dimensions need to be available in staging area.	All data will be available because Extract and load occur in one single action.

Parameters	ETL	ELT
Aggregations	Complexity increase with the additional amount of data in the dataset.	Power of the target platform can process significant amount of data quickly.
Calculations	Overwrites existing column or Need to append the dataset and push to the target platform.	Easily add the calculated column to the existing table.
Maturity	The process is used for over two decades. It is well documented and best practices easily available.	Relatively new concept and complex to implement.
Hardware	Most tools have unique hardware requirements that are expensive.	Being Saas hardware cost is not an issue.
Support for Unstructured Data	Mostly supports relational data	Support for unstructured data readily available.

THANK YOU