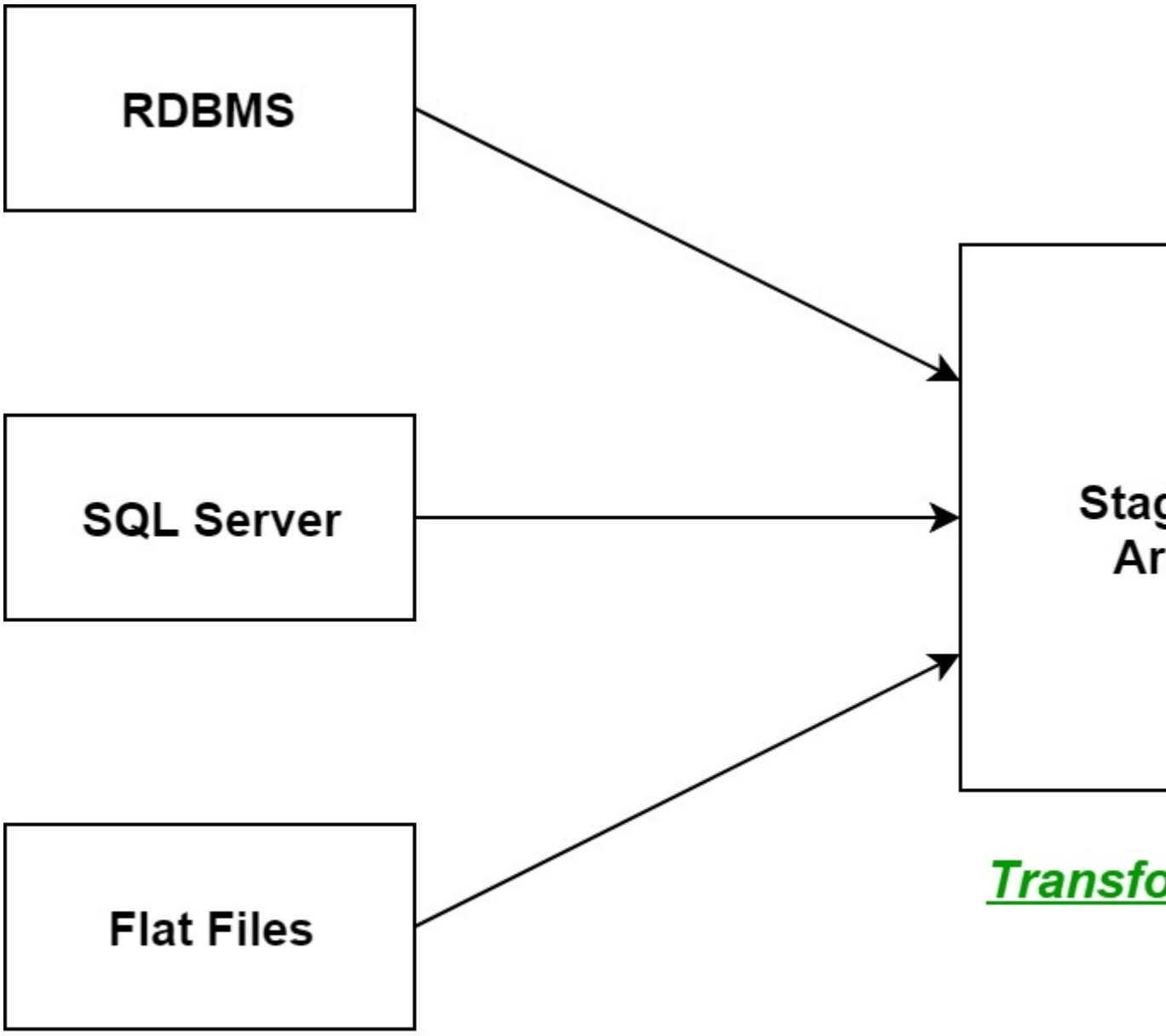
ETL Process in Data Warehouse

INTRODUCTION:

1. ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse. The process of ETL can be broken down into the following three stages:
2. **Extract**: The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area.
3. **Transform**: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.
4. **Load**: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.
5. The ETL process is an iterative process that is repeated as new data is added to the warehouse. The process is important because it ensures that the data in the data warehouse is accurate, complete, and up-to-date. It also helps to ensure that the data is in the format required for data mining and reporting.

Additionally, there are many different ETL tools and technologies available, such as Informatica, Talend, DataStage, and others, that can automate and simplify the ETL process.

ETL is a process in Data Warehousing and it stands for **Extract**, **Transform** and **Load**. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area, and then finally, loads it into the Data Warehouse system.



Extraction

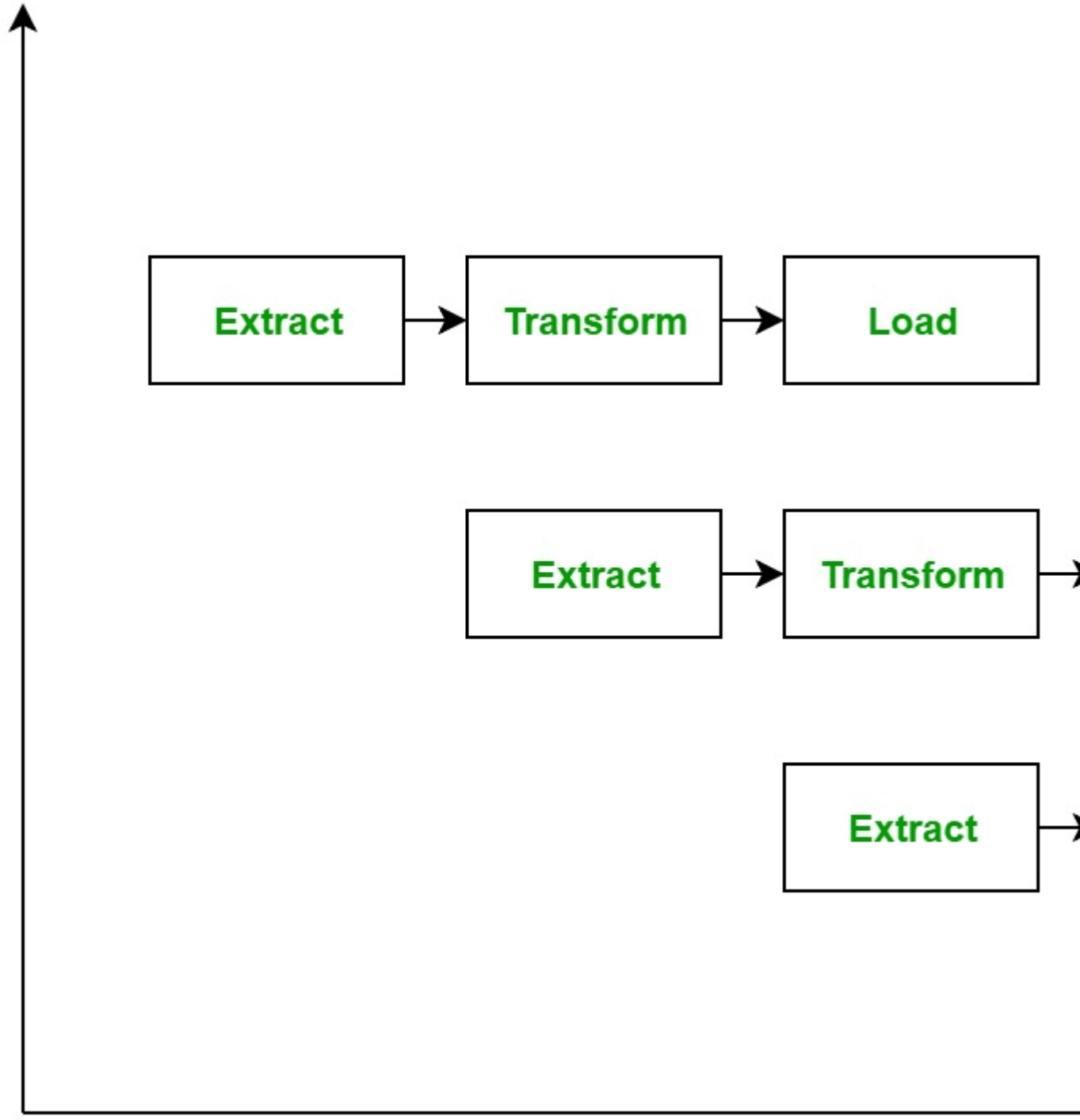
Let us understand each step of the ETL process in-depth:

1. **Extraction:**
   The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML, and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also.

Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.

2. **Transformation:**
   The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:
   - Filtering – loading only certain attributes into the data warehouse.
   - Cleaning – filling up the NULL values with some default values, mapping U.S.A, United States, and America into USA, etc.
   - Joining – joining multiple attributes into one.
   - Splitting – splitting a single attribute into multiple attributes.
   - Sorting – sorting tuples on the basis of some attribute (generally key-attribute).
3. **Loading:**
   The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

ETL process can also use the pipelining concept i.e. as soon as some data is extracted, it can transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed. The block diagram of the pipelining of ETL process is shown below:



ETL Tools: Most commonly used ETL tools are **Hevo**, Sybase, Oracle Warehouse builder, CloverETL, and MarkLogic.

Data Warehouses: Most commonly used Data Warehouses are **Snowflake**, Redshift, BigQuery, and Firebolt.

ADVANTAGES OR DISADVANTAGES:

Advantages of ETL process in data warehousing:

1. **Improved data quality:** ETL process ensures that the data in the data warehouse is accurate, complete, and up-to-date.
2. **Better data integration:** ETL process helps to integrate data from multiple sources and systems, making it more accessible and usable.
3. **Increased data security:** ETL process can help to improve data security by controlling access to the data warehouse and ensuring that only authorized users can access the data.
4. **Improved scalability: E**TL process can help to improve scalability by providing a way to manage and analyze large amounts of data.
5. **Increased automation:** ETL tools and technologies can automate and simplify the ETL process, reducing the time and effort required to load and update data in the warehouse.

Disadvantages of ETL process in data warehousing:

1. **High cost:** ETL process can be expensive to implement and maintain, especially for organizations with limited resources.
2. C**omplexity:** ETL process can be complex and difficult to implement, especially for organizations that lack the necessary expertise or resources.
3. **Limited flexibility:** ETL process can be limited in terms of flexibility, as it may not be able to handle unstructured data or real-time data streams.
4. **Limited scalability**: ETL process can be limited in terms of scalability, as it may not be able to handle very large amounts of data.
5. **Data privacy concerns**: ETL process can raise concerns about data privacy, as large amounts of data are collected, stored, and analyzed.

Overall, ETL process is an essential process in data warehousing that helps to ensure that the data in the data warehouse is accurate, complete, and up-to-date. However, it also comes with its own set of challenges and limitations, and organizations need to carefully consider the costs and benefits before implementing them.