ETL Process in Data Warehouse

INTRODUCTION:

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:

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.

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.**

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.**

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.

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

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.

Transformation:

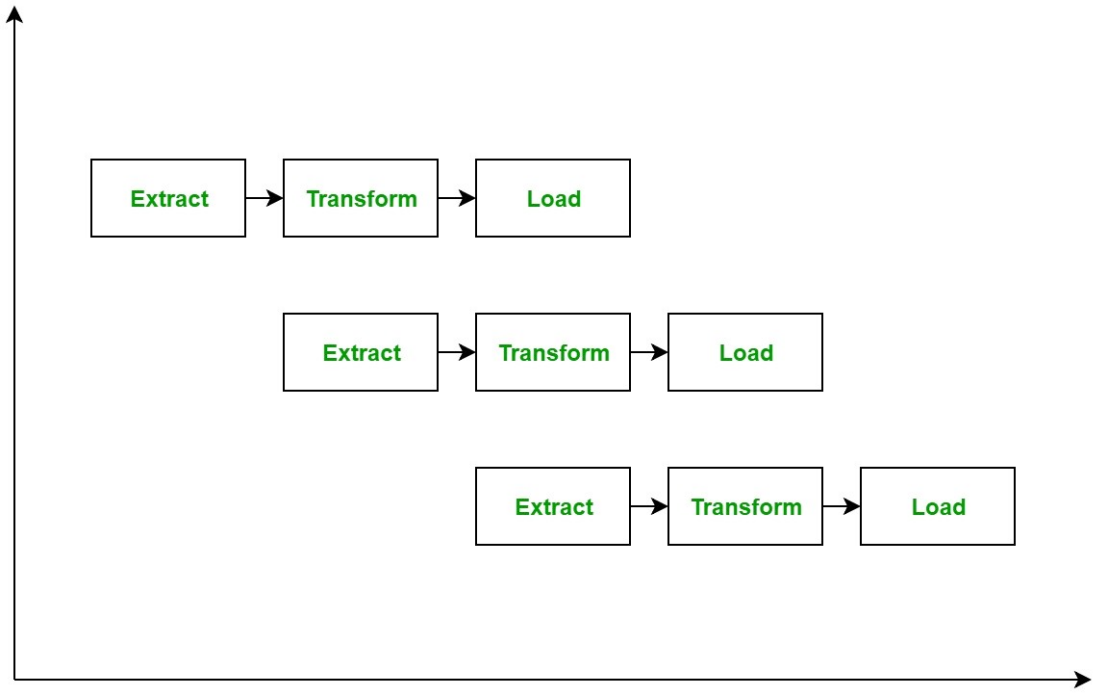
The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:

- **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).**

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:



CloverETL, and MarkLogic.

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

ETL Tools: Most commonly used ETL tools are Hevo, Sybase, Oracle Warehouse builder,

1. Data Extraction (E): Extract data from various source systems. This could include databases, flat files, APIs, and other data repositories. Db2 Warehouse supports various

data extraction methods:

- Use SQL to extract data from relational databases.
- Use ETL tools like Apache Nifi, Apache Spark, or Talend for structured data extraction.
- Utilize APIs for extracting data from web services.

- Implement custom scripts to extract data from less structured sources, such as log files.
- 2. Data Transformation (T): Once data is extracted, it often needs to be transformed to fit the target data model and to ensure data quality. Transformation can include tasks like:
- Data cleansing to remove duplicates, handle missing values, and correct errors.
- Data enrichment by merging data from multiple sources.
- Data aggregation and summarization.
- Applying business rules to standardize data.

3. Data Loading (L): Load the transformed data into Db2 Warehouse. This can be done through various methods:

- Bulk loading using tools like IBM Data Movement Tool, Apache Nifi, or Db2's data import utilities.
- Real-time streaming for continuous data updates.
- Incremental loading for data that changes over time.

4. Data Modeling: Design and implement a data model within Db2 Warehouse that aligns with your organization's reporting and analytical needs. This involves creating tables, views, and relationships that facilitate easy data retrieval.

5. SQL Query and Analysis:

- Data Exploration: Enable data architects to explore data using SQL queries and various analysis techniques. They can connect to Db2 Warehouse using SQL clients, such as IBM Data Studio, DBeaver, or even command-line tools.
- OLAP and Data Warehousing Techniques: Utilize SQL for OLAP (Online Analytical Processing) operations. Data architects can create cubes, use

window functions, and build complex queries to support multidimensional analysis.

- Machine Learning Integration: Enable data scientists and analysts to build and deploy machine learning models within the data warehouse to gain insights and predictions from the data.

Reporting and Visualization: Integrate reporting and visualization tools like Tableau, Power BI, or IBM Cognos with Db2 Warehouse to create interactive dashboards and reports.

6. Data Security: Implement role-based access control and encryption mechanisms to secure sensitive data in the data warehouse.

7. Data Governance: Establish data governance practices to ensure data quality, compliance, and data lineage.

8. Monitoring and Maintenance: Set up monitoring and alerting systems to keep an eye on the health of the data warehouse. Regularly maintain and optimize the database for performance.

9. Documentation: Document the ETL processes, data models, and best practices to ensure knowledge transfer and future scalability.

10. Training: Provide training to data architects, analysts, and other relevant staff to ensure they can effectively work with Db2 Warehouse.

11. Automation: Implement automation for ETL processes, monitoring, and maintenance tasks to reduce manual effort and ensure data consistency.

12. Scalability: Plan for scalability and growth, considering factors like increasing data volume and user demands.

Remember that building and maintaining a data warehouse is an ongoing process, and it's important to continually assess and adapt to changing business needs and data sources.

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.