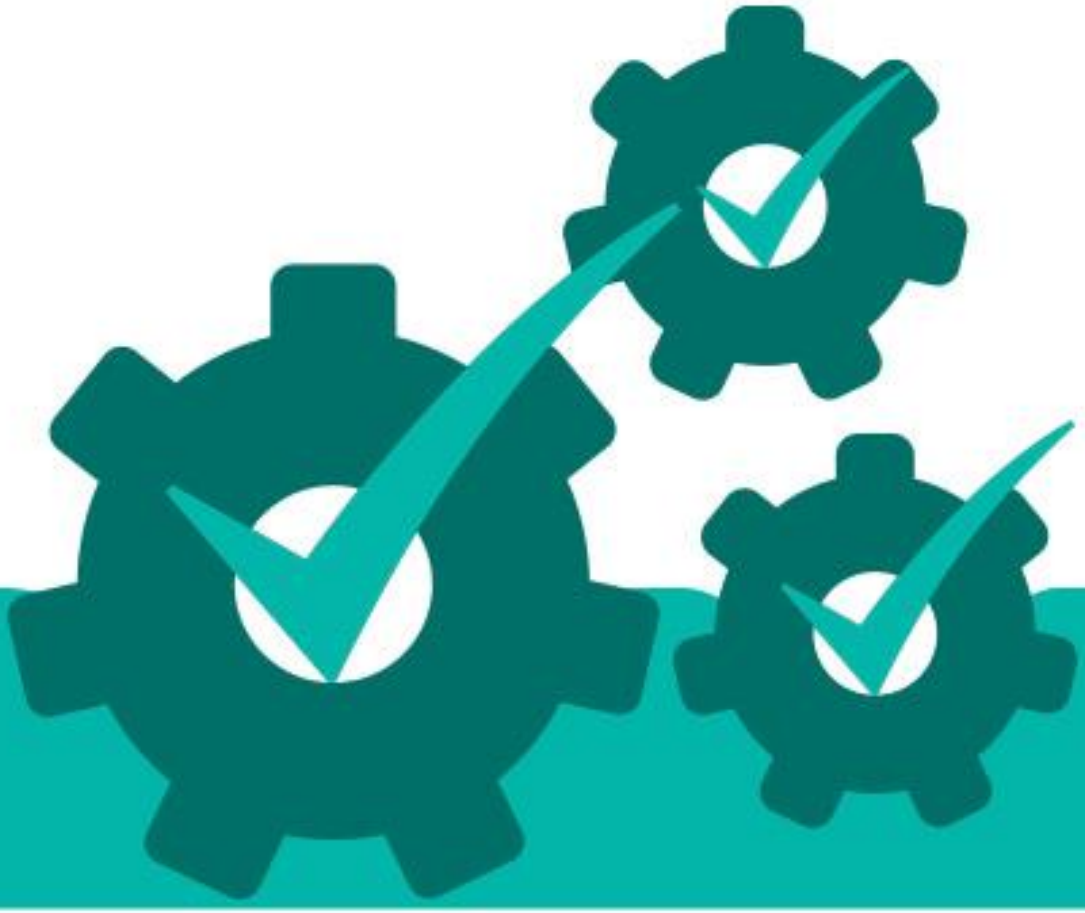
ETL Testing



tutorialspoint

SIMPLYEASYLEARNING

www.tutorialspoint.com

https://www.facebook.com/tutorialspointindia

https://twitter.com/tutorialspoint

1. ETL – Introduction

The data in a Data Warehouse system is loaded with an ETL (Extract, Transform, Load) tool. As the name suggests, it performs the following three operations:

- Extracts the data from your transactional system which can be an Oracle, Microsoft, or any other relational database,
- Transforms the data by performing data cleansing operations, and then
- Loads the data into the OLAP data Warehouse.

You can also extract data from flat files like spreadsheets and CSV files using an ETL tool and load it into an OLAP data warehouse for data analysis and reporting. Let us take an example to understand it better.

Example

Let us assume there is a manufacturing company having multiple departments such as sales, HR, Material Management, EWM, etc. All these departments have separate databases which they use to maintain information w.r.t. their work and each database has a different technology, landscape, table names, columns, etc. Now, if the company wants to analyze historical data and generate reports, all the data from these data sources should be extracted and loaded into a Data Warehouse to save it for analytical work.

An ETL tool extracts the data from all these heterogeneous data sources, transforms the data (like applying calculations, joining fields, keys, removing incorrect data fields, etc.), and loads it into a Data Warehouse. Later, you can use various Business Intelligence (BI) tools to generate meaningful reports, dashboards, and visualizations using this data.

Difference between ETL and BI Tools

An ETL tool is used to extract data from different data sources, transform the data, and load it into a DW system; however a BI tool is used to generate interactive and ad-hoc reports for end-users, dashboard for senior management, data visualizations for monthly, quarterly, and annual board meetings.

The most common ETL tools include: SAP BO Data Services (BODS), Informatica – Power Center, Microsoft – SSIS, Oracle Data Integrator ODI, Talend Open Studio, Clover ETL Open source, etc.

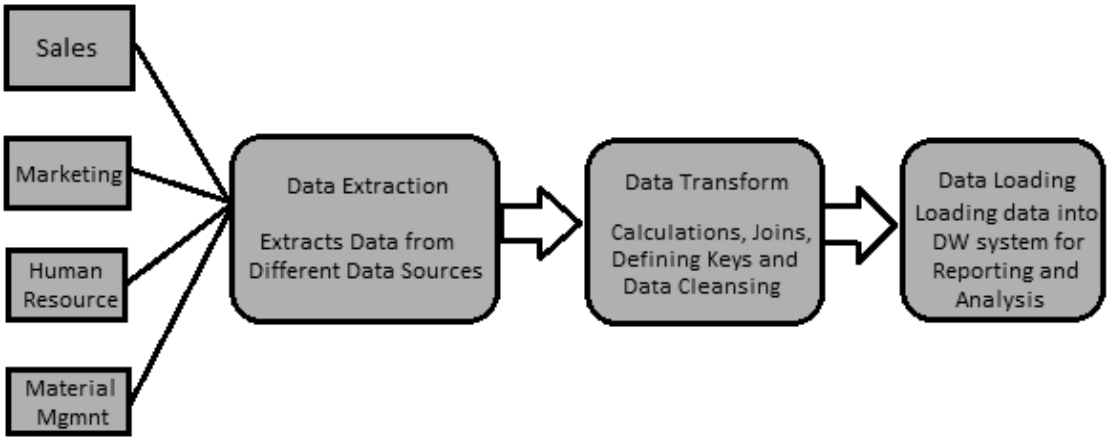
Some popular BI tools include: SAP Business Objects, SAP Lumira, IBM Cognos, JasperSoft, Microsoft BI Platform, Tableau, Oracle Business Intelligence Enterprise Edition, etc.

ETL Process

Let us now discuss in a little more detail the key steps involved in an ETL procedure –

Extracting the Data

It involves extracting the data from different heterogeneous data sources. Data extraction from a transactional system varies as per the requirement and the ETL tool in use. It is normally done by running scheduled jobs in off-business hours like running jobs at night or over the weekend.



Transforming the Data

It involves transforming the data into a suitable format that can be easily loaded into a DW system. Data transformation involves applying calculations, joins, and defining primary and foreign keys on the data. For example, if you want % of total revenue which is not in database, you will apply % formula in transformation and load the data. Similarly, if you have the first name and the last name of users in different columns, then you can apply a concatenate operation before loading the data. Some data doesn't require any transformation; such data is known as **direct move** or **pass through data**.

Data transformation also involves data correction and cleansing of data, removing incorrect data, incomplete data formation, and fixing data errors. It also includes data integrity and formatting incompatible data before loading it into a DW system.

Loading the Data into a DW System

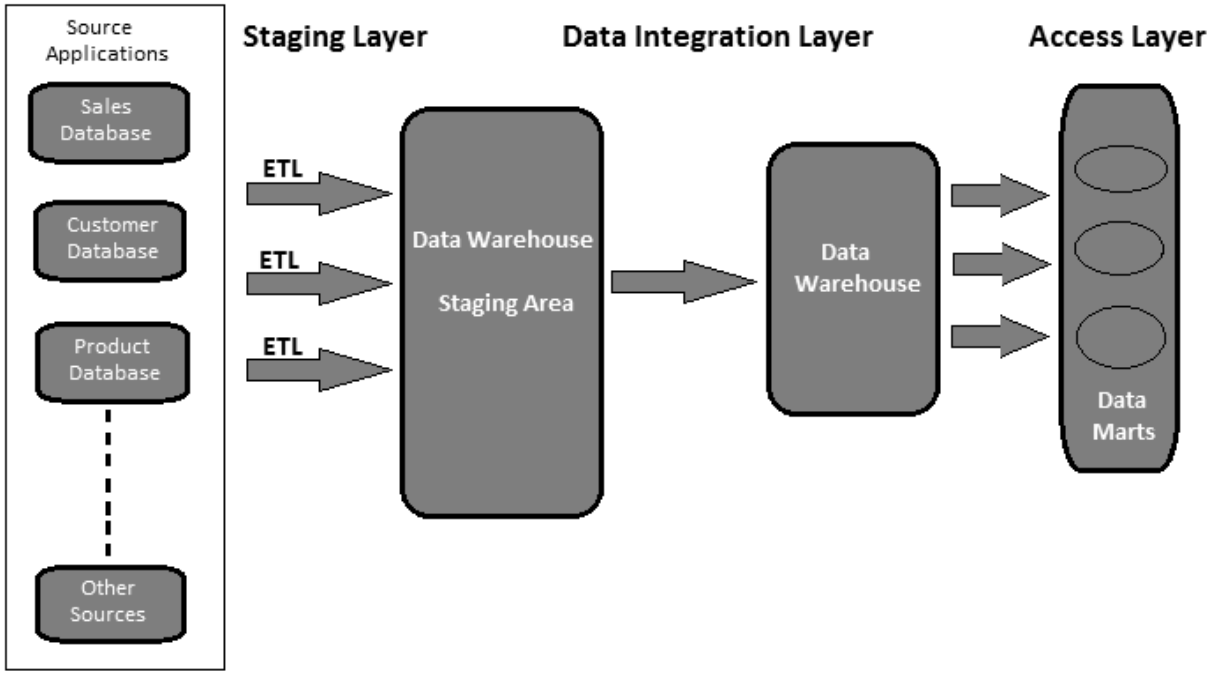
It involves loading the data into a DW system for analytical reporting and information. The target system can be a simple delimited flat file or a data warehouse.

ETL Tool Function

A typical ETL tool-based data warehouse uses staging area, data integration, and access layers to perform its functions. It's normally a 3-layer architecture.

- **Staging Layer** – The staging layer or staging database is used to store the data extracted from different source data systems.
- **Data Integration Layer** – The integration layer transforms the data from the staging layer and moves the data to a database, where the data is arranged into hierarchical groups, often called **dimensions**, and into **facts** and **aggregate facts**. The combination of facts and dimensions tables in a DW system is called a **schema**.
- **Access Layer** – The access layer is used by end-users to retrieve the data for analytical reporting and information.

The following illustration shows how the three layers interact with each other.



2. ETL Testing – Tasks

ETL testing is done before data is moved into a production data warehouse system. It is sometimes also called as **table balancing** or **production reconciliation**. It is different from database testing in terms of its scope and the steps to be taken to complete this.

The main objective of ETL testing is to identify and mitigate data defects and general errors that occur prior to processing of data for analytical reporting.

ETL Testing – Tasks to be Performed

Here is a list of the common tasks involved in ETL Testing –

1. Understand the data to be used for reporting
2. Review the Data Model
3. Source to target mapping
4. Data checks on source data
5. Packages and schema validation
6. Data verification in the target system
7. Verification of data transformation calculations and aggregation rules
8. Sample data comparison between the source and the target system
9. Data integrity and quality checks in the target system
10. Performance testing on data

3. ETL vs Database Testing

Both ETL testing and database testing involve data validation, but they are not the same. ETL testing is normally performed on data in a data warehouse system, whereas database testing is commonly performed on transactional systems where the data comes from different applications into the transactional database.

Here, we have highlighted the major differences between ETL testing and Database testing.

ETL Testing

ETL testing involves the following operations:

1. Validation of data movement from the source to the target system.
2. Verification of data count in the source and the target system.
3. Verifying data extraction, transformation as per requirement and expectation.
4. Verifying if table relations – joins and keys – are preserved during the transformation.

Common ETL testing tools include **QuerySurge**, **Informatica**, etc.

Database Testing

Database testing stresses more on data accuracy, correctness of data and valid values. It involves the following operations:

1. Verifying if primary and foreign keys are maintained.
2. Verifying if the columns in a table have valid data values.
3. Verifying data accuracy in columns. **Example**: Number of months column shouldn't have a value greater than 12.
4. Verifying missing data in columns. Check if there are null columns which actually should have a valid value.

Common database testing tools include **Selenium**, **QTP**, etc.