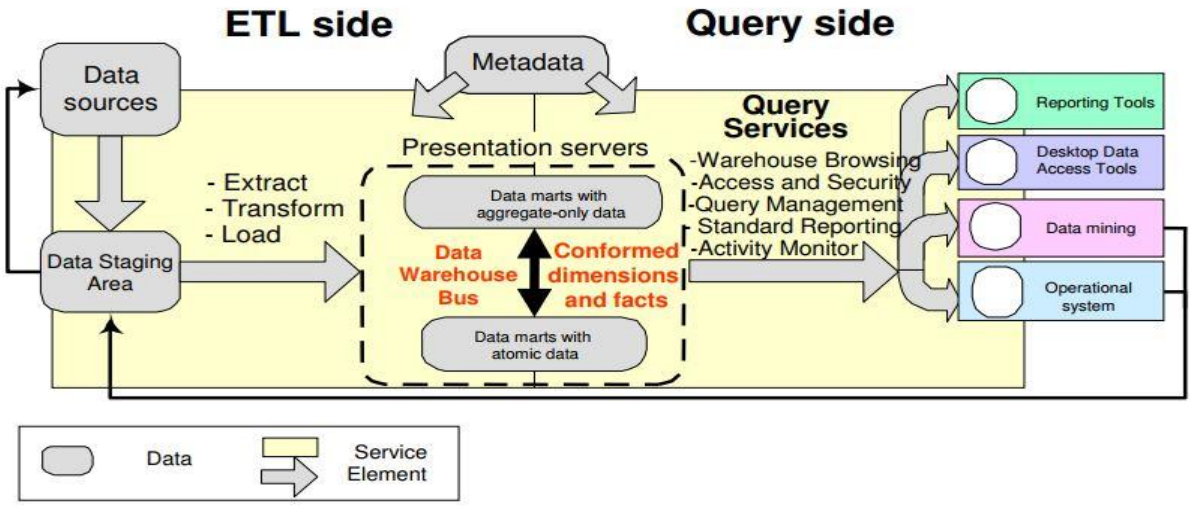
ETL

(Extract, Transform & Load)

Extract, transform, and load (ETL) is the process of integrating data from multiple, typically disparate, sources and bringing them together into one central location. It is a key component to businesses successfully making use of data in a data warehouse.

Sure, the process itself is fairly straightforward, and when done right, ETL prepares an organization for powerful business intelligence initiatives. However, a lot goes into a successful ETL process. Let's discuss the three steps involved and why data management practices are an essential foundation to carrying ETL out properly.



ETL Process :-

If you'll remember, ETL stands for Extract, Transform, and Load. An organization looking to complete an ETL process must:

1) Extract data from the database(s) to integrate it from various systems or applications

2) Transform data so that it matches the target system's required formatting, and

3) Load the final data into the target system.

Extraction

Extraction is the action of extracting data from a source system to be processed at a later stage. This step is focused on obtaining data as efficiently and with as little impact to the source system as possible.

Businesses today collect and store data in a variety of sources—each with their own way of organizing and formatting that data—and oftentimes the sheer volume of data can make this first step in the ETL process the most time-consuming.

Transformation

After the desired data has been extracted, it then undergoes a transformation (i.e. conversion) to meet the requirements of the target system. This step can involve:

- Cleanising and validating data to ensure only quality data is migrated to the new system
- Sorting the data into columns or rows to improve usability and searchability
- Combining or merging data from multiple source systems and duplicating that data
- Applying business rules to data
- Creating data validation rules that can be automated to check for data for quality and accuracy

This process entails several transformation types that ensure the quality and integrity of data. Without this step, businesses can't be confident in the data being migrated or integrated into the target system—which can mean weeks or even months of effort and budget lost!

Load

The Load step concludes the ETL process with the loading of the extracted and transformed data into the end system. The successful completion as well as complexity of this step is dependent on the volume of data, structure of that data, and frequency at which you will load that data.

ETL Tools :-

List of the most popular ETL tools:

- Informatica - Power Center
- IBM - Websphere DataStage(Formerly known as Ascential DataStage)
- SAP - BusinessObjects Data Integrator
- IBM - Cognos Data Manager (Formerly known as Cognos DecisionStream)
- Microsoft - SQL Server Integration Services
- Oracle - Data Integrator (Formerly known as Sunopsis Data Conductor)
- SAS - Data Integration Studio
- Oracle - Warehouse Builder
- AB Initio
- Information Builders - Data Migrator
- Pentaho - Pentaho Data Integration

- Embarcadero Technologies - DT/Studio
- IKAN - ETL4ALL
- IBM - DB2 Warehouse Edition
- Pervasive - Data Integrator
- ETL Solutions Ltd. - Transformation Manager
- Group 1 Software (Sagent) - DataFlow
- Sybase - Data Integrated Suite ETL
- Talend - Talend Open Studio
- Expressor Software - Expressor Semantic Data Integration System
- Elixir - Elixir Repertoire
- OpenSys – CloverETL

Pentaho :-

Pentaho Data Integration (PDI, also called *Kettle*) is the component of Pentaho responsible for the Extract, Transform and Load (ETL) processes. Though ETL tools are most frequently used in data warehouses environments, PDI can also be used for other purposes:

- Migrating data between applications or databases
- Exporting data from databases to flat files
- Loading data massively into databases
- Data cleansing
- Integrating applications

PDI is easy to use. Every process is created with a graphical tool where you specify what to do without writing code to indicate how to do it; because of this, you could say that PDI is *metadata oriented*.

PDI can be used as a standalone application, or it can be used as part of the larger Pentaho Suite. As an ETL tool, it is the most popular open source tool available. PDI supports a vast array of input and output formats, including text files, data sheets, and commercial and free database engines. Moreover, the transformation capabilities of PDI allow you to manipulate data with very few limitations.

Spoon :-

Spoon is a graphical user interface that allows you to design transformations and jobs that can be run with the Kettle tools — Pan and Kitchen. Pan is a data transformation engine that performs a multitude of functions such as reading, manipulating, and writing data to and from various data sources. Kitchen is a program that executes jobs designed by Spoon in XML or in a database repository. Jobs are usually scheduled in batch mode to be run automatically at regular intervals.

Kettle is an acronym for "Kettle E.T.T.L. Environment." Kettle is designed to help you with your ETTL needs, which include the Extraction, Transformation, Transportation and Loading of data.

Transformations and Jobs can describe themselves using an XML file or can be put in a Kettle database repository. Pan or Kitchen can then read the data to execute the steps described in the transformation or to run the job. In summary, Pentaho Data Integration makes data warehouses easier to build, update, and maintain.