## ETL -Extract, transform, load

* ETL process is used in data warehouse.
* When we say ETL that means we extract the data on the basis of the output transform it as per requirements and then we load it into the datawarehouse using different analytical tools.
  + For instance we have the raw, next what we do is we extract only those informations’ which we currently find important for the storage and different analysis.
* We extract the data from different sources :
  + for example we have 3 warehouses for the ecommerce, so we collect the data from all the three sources …that is being extracting the data.
* We transform the data using various techniques:
  + Cleaning
  + Preprocessing(removing missing values , rectifying the formats i.e the whole column should maintain the same format eg: date )
  + Structuring data in tabular form
  + Normalizing the data i.e no redundancies, transactional dependencies,etc
  + Formatting the structure as per the datawarehouse format
  + Stored in flat-files or in schema form
  + Sorting and many more techniques
  + Here the data is in the atomic form i.e it cannot be divided further into any dimensions or attributes
* Moreover, we have loading,
  + In loading we load the data in the form of data marts
  + From which the user can easily access the data using insert different queries
  + For example, as we have taken it earlier :
    - We have 3 warehouses so we divided data in transform process into the various databases, so what we do here is we load these data into different datamarts say order is useful to marketing analysis datamart as well as in finance data mart
    - Similarly product is useful for marketing as well as finance data mart.
    - We load data by either **refresh** i.e twh whole datawarehouse is emptied and then data is loaded or through **update** i.e we just don’t hamper the existing data and we just load the new data to it
  + So this is how it works in ETL .
* Advantages:
  + Faster execution time
  + Targeted data to retrieve or fetch.
* Disadvantages:
  + Flexibility : as stated earlier if we need another data or column for the raw data we have to rectify the whole process of ETL
* Now ETL is basically where the whole process runs simultaneously
  + For example if the data is to be processed on every Wednesday at 9
  + So what will happen is, the data is to be transformed and loaded simultaneously that is if the data is not transformed it won’t be loaded
  + So it becomes the drawback for the same.
  + Because it may take the days of time and until that it won’t be processed
* SO TO OVERCOME THIS WE HAVE “ ELT ”

## ETL -Extract, load, transform

* Here we extract , load and then transform
* So refering to the above example if we want to process the data on Wednesday at 9 so what it will do is ….the data will be extracted and loaded to the datawarehouse and it will be transformed when the query is run
* So loading and transforming both are independent of each other unlike ETL process.

Google , amazon and various company uses ETL, ELT or say hybrid as per the requirement of the data

* So for example amazon have data pipeline called AWS Glue inside which the ETL process is done and it uses datawarehouse called Amazon Redshift
* So what is data pipeline?
  + It is basically through which the data is inserted from one end and it goes through various processes and not only ETL (specific type in data pipeline)and the processed data comes out of the other end.
* Similarly for Microsoft the datawarehouse is AZURE Synapse and data pipeline is AZURE data factory.
* For google the datawarehouse is BigQuery and datapipeline is Cloud Dataflow.