# Week 8 Contribution - Tanmay Munigala – 12742995

## ETL Pipeline Notes

### What is ETL?

ETL is a data integration process, which the procedure of copying data from one or more sources into a destination system which represents the data differently in a different context than the original source.

It stands for Extract, Transform and Load, which are the 3 main steps in the process.

#### Extract

In this phase, the data is going to be extracted from any number of source systems. This is the main aspect of ETL as it is the stage-setter for the rest of the steps. The main things to keep in mind during this phase is that different sources may store data is different formats or different systems.

Part of the extract phase is also data validation, to confirm whether the data pulled has the correct values. If the validation is fails, the data should be rejected.

#### Transform

In this phase, the collected data from the Extract phase is worked upon to convert it into the data format or in the proper form to the target source. A set of rules/functions are used on this step for the transformation process to take place.

#### Load

In the phase of the process, the transformed data from the Transform phase is loaded onto the target system or data store. The main concept to keep in mind here is that the target system may have a different database schema that the source and will have to be transformed and loaded accordingly. Based on business requirements, sometimes the load step involves overwritten any existing data in the target data store with newer records.

Some tools to help us with the ETL process in the context of our Node JS application:

1. HeyNode - <https://heynode.com/>

2. Empujar - <https://github.com/taskrabbit/empujar>

3. Nextract - <https://github.com/cauld/nextract>