

# ETL

▼ Class	07_pipeline
🕒 Created	@Jan 13, 2021 10:31 AM
🔗 Materials	
☑ Reviewed	<input type="checkbox"/>
▼ Type	

## ETL - Extract Transform Load

### 1) What is ETL?

Extract - Transform - Load

- Extract data from one or more data sources. The data sources can generally be of any kind and the number is (theoretically) not limited
- Transform the data. The easiest transformation is to leave the data as it is. A transformation could be filtering of data, aggregation of data, performing calculations (implementing business logic), changing data types, ...
- Load the transformed data into a new form of Data Storage; usually this will be some kind of database

### 2) Why would we implement an ETL process?

- Source systems (source databases) will be customized for different use cases. In larger companies you will have multiple source systems.
- Focus of the source systems is (OLTP - Online Transaction Platform):
  - to provide persistent storage
  - to manage data quality
  - designed so they are efficiently updating and inserting information

- If we consider requirements from the perspective of a Data Analyst / Data Scientist the focus will shift:
  - Query a lot of data across tables → joins are computationally expensive
  - Focus on fast querying and less on updating and inserting data

2 main reasons why ETL processes are so important:

- Combine multiple data sources
- Fulfill requirements of Analytical applications