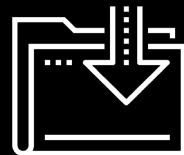




Project #2: Extract, Transform, and Load

Data Boot Camp
Lesson 13.1







The Week Ahead!

▼ Day 1 (today):

/ Introduction to the ETL process: working through activities

▼ Introduction to the ETL project

/ Goals

/ Requirements

▼ Working toward a feasible project idea with:

/ Instructors

/ TAs

▶ Submit project proposal

▼ Day 2:

/ Working on projects

/ Full assistance from instructors and TAs

▼ Day 3:

/ Project due date: presentation!

/ Discussion



Case Study Project Requirements

Data sources

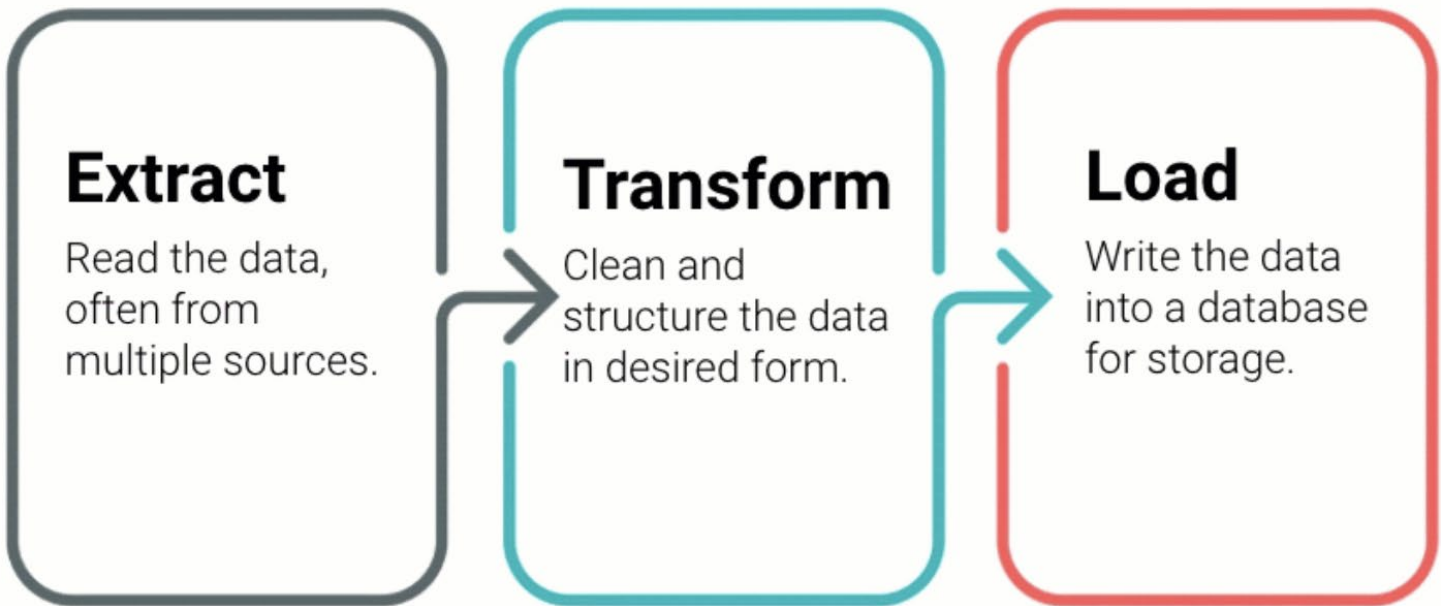
- You must have two (minimum) or more sources.
- Recommended sources:
 - Kaggle
 - Data.world
 - Google Dataset Search (<https://datasetsearch.research.google.com/>)
 - APIs may be used as an alternative source
- Once your datasets are identified, perform ETL and create documentation.
 - Documentation must have:
 - Datasets used and their sources
 - Types of data wrangling performed (data cleaning, joining, filtering, and aggregating)
 - The schemata used in the final production database



Introduction to ETL

ETL

Data integration is an important part of working with data.





Introduction to ETL

Extract

Data may come from disparate sources, such as:

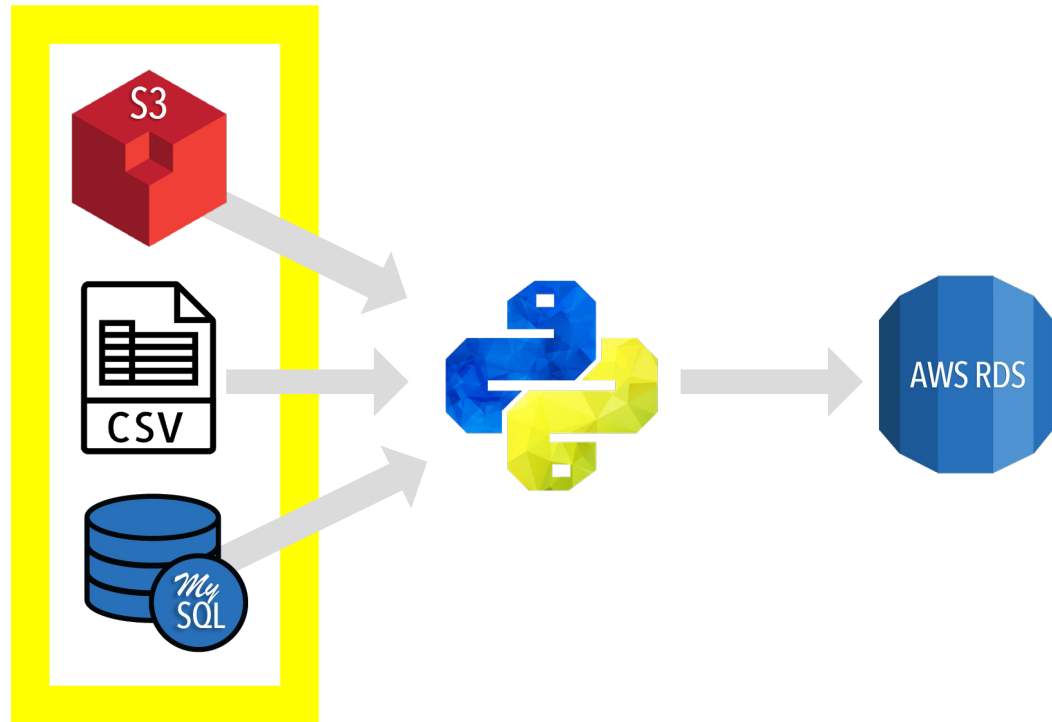
CSV files

JSON files

HTML tables

SQL databases

Spreadsheets



Extract



Introduction to ETL

Transform

Transform the data to suit business needs.
This may include:

Data cleaning

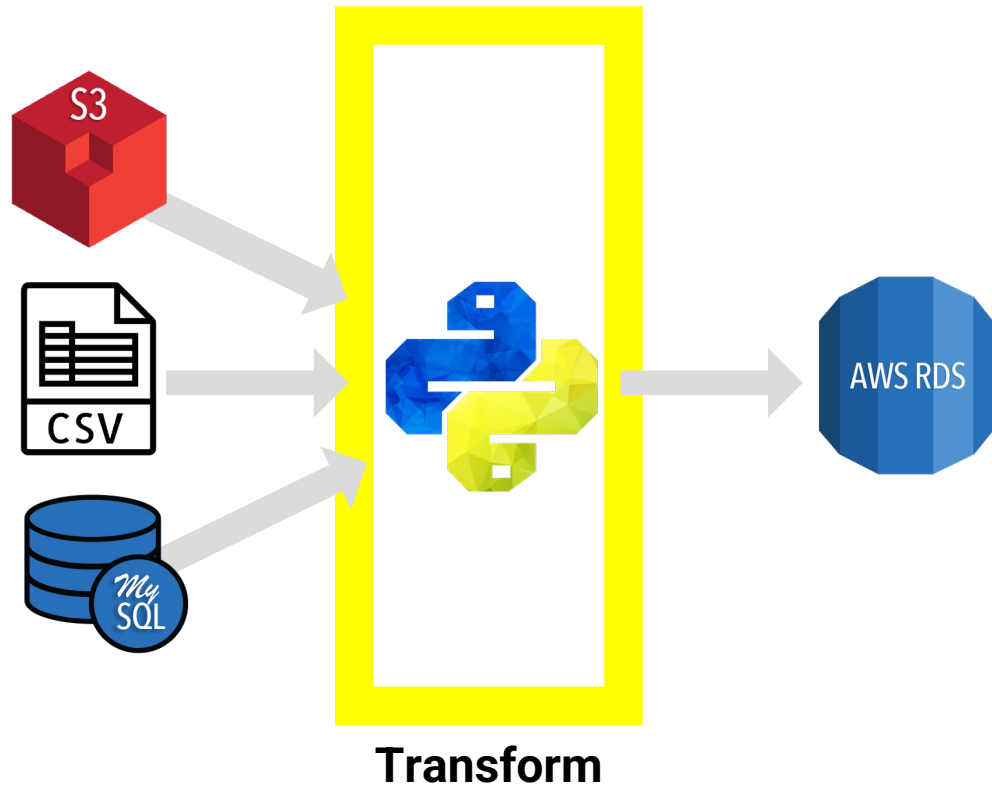
Summarization

Selection

Joining

Filtering

Aggregating





Note: We will use Python and Pandas for transformation, which can also be done with SQL or a specialized ETL tool.



Introduction to ETL

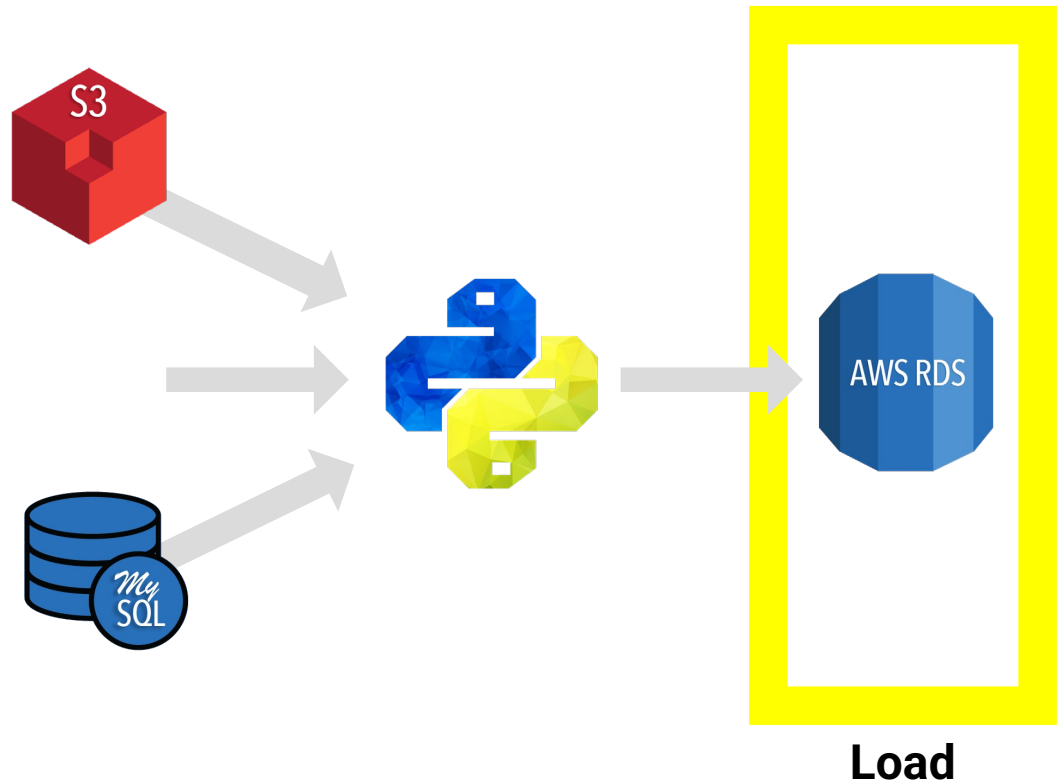
Load

Load the data into a final database that can be used for future analysis or business use.

Can be a relational or non-relational database

Can be local or in the cloud

Can be a data lake or data warehouse





ETL with Pandas

- Not limited to **Pandas**, the **ETL** process is performed with a variety of tools and file formats.
- For this demonstration, we will use the following:



 pandas



PostgreSQL

Project Overview

Project Week! (This Week)

Day 1:



Form groups



Identify datasets



Begin ETL on the data

Day 2:

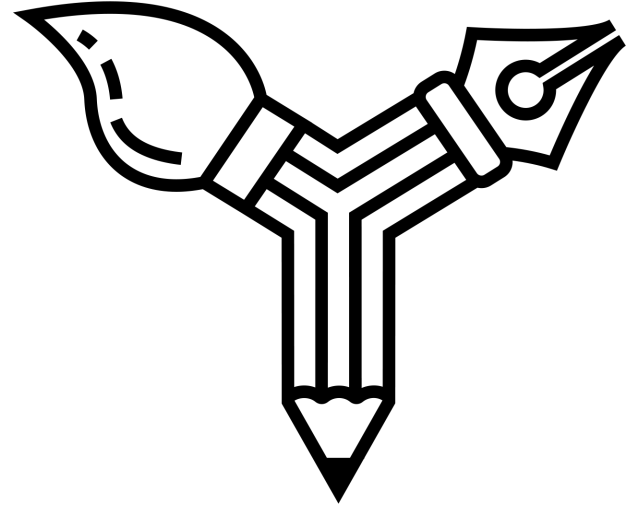


Database development

Day 3:



Complete final report





Project Proposals



Team effort

Due to the short timeline, teamwork will be crucial to the success of this project! Work closely with your team through all phases of the project to ensure that there are no surprises at the end of the week.

Working in a group enables you to tackle more difficult problems than you'd be able to work on alone. In other words, working in a group allows you to work smart and dream big. Take advantage of it!



Project proposal

Before you start writing any code, remember that you only have one week to complete this project. View this project as a typical assignment from work. Imagine that a bunch of data came in and you and your team are tasked with migrating it to a production database.

Take advantage of your instructor and TA support during office hours and class project work time. They are a valuable resource and can help you stay on track.

Project 2: ETL

Data Cleanup and Analysis Requirements

You will also be responsible for:



Citing the data sources from which you will extract.



Extracting the data from their existing locations.



Transforming the data (i.e. cleaning, joining, filtering, aggregating, etc).



Loading the data to a database (relational or non-relational).

Report Requirements

You will also be responsible for preparing a formal report that covers:



Extract: your original data sources and how the data were formatted (CSV, JSON, pgAdmin 4, etc.).



Transform: what data cleaning or transformation was required.



Load: the final database, tables/collections, and why this was chosen.

Project Rubric

Rubric at a Glance

Categories for grading



Project proposal (20 points)



Technical report (20 points)



GitHub repository (20 points)

Data Suggestions

Data Suggestions

Feel free to ask us (the instructional staff) for input, but our general advice is to stick to data sources that:



Are sufficiently large.



Have a consistent format.



Ideally, contain more data than needed.



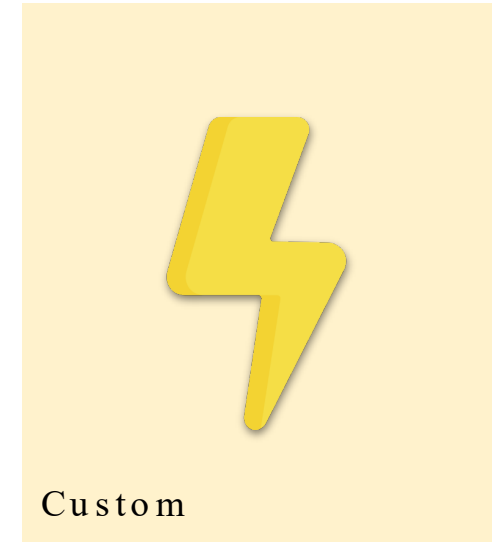
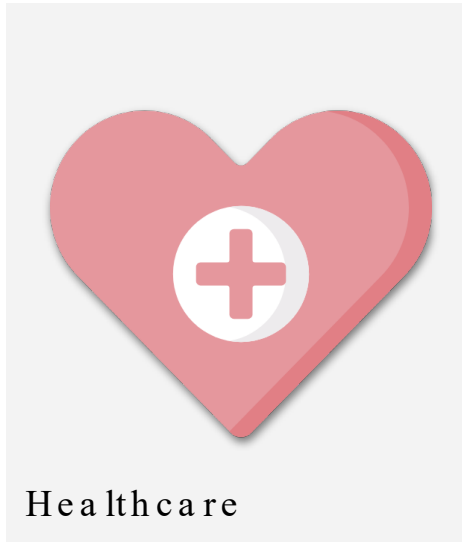
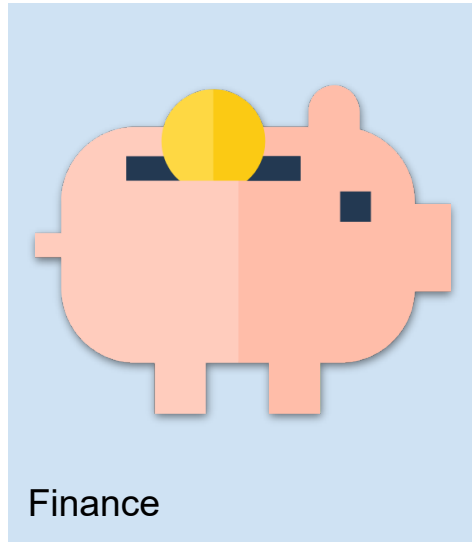
Are well documented.

Choosing a Project Track

Choosing a Project Track

This project gives you the ability to focus your efforts within a specific industry.

Here are the specializations:





ETL and Finance



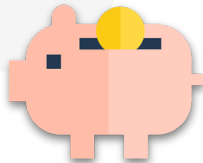
When to use ETL in finance

Current treasury benchmarks are at an all-time low, and a financial analyst has decided to study the last 30 years worth of rates.

After pulling historic data, they clean and explore them to perform analysis, with the intent of predicting future benchmark trends.

Once the history data have been collected, processed, and loaded into a database, the financial analyst turns their attention to current data. Using an API, they pull the most up-to-date information with the intent of adding it to their established database.

They've already extracted the new data, but before loading them into the existing database, they need to ensure that they have the correct format. Once the data are transformed, they can load them to the database and continue on with the analysis.



ETL and Healthcare



When to use ETL in healthcare

An analyst working at a major hospital is tasked with reviewing policies regarding the upcoming flu season. They're keeping the following questions in mind:

- How many patients does the hospital expect this year?
- How severe will the flu season be this year?
- Will there be regional differences? Similarities?

The analyst wants to collect data from different sources to analyze them and predict the future flu season.

Before combining the hospital's own data with regional data acquired elsewhere, the analyst will need to extract the new data, transform them to match the existing data, and then load them into the database.





ETL in the Wild



Customize it!

There are several other industries utilizing the ETL process as well.

- In marketing, analysts may acquire data from competitors to see how their product fares. Multiple data sources would need to be extracted, transformed, and loaded into a common database prior to analysis.
- An analyst working for a large retail chain finds themselves in charge of moving a legacy database into a cloud-based data warehouse.
- An entrepreneur has a big business idea but wants to get a feel for their product idea. They use web scraping and APIs to pull data from a variety of social media platforms with the intent of analyzing consumer reactions.





Questions?



Time to divide into teams!
Good luck!

