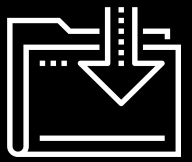
ETL: Extract, Transform, Load

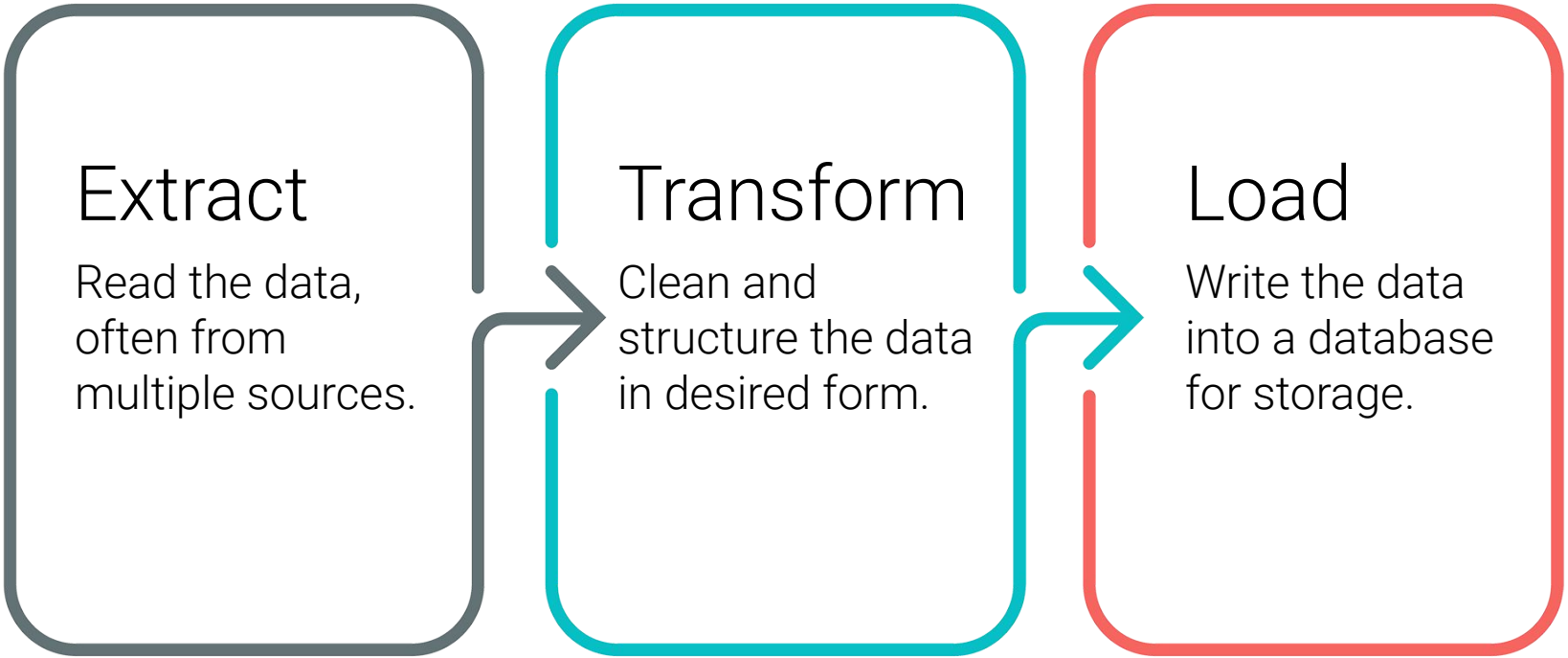
Data Boot Camp

Lesson 13.1



ETL

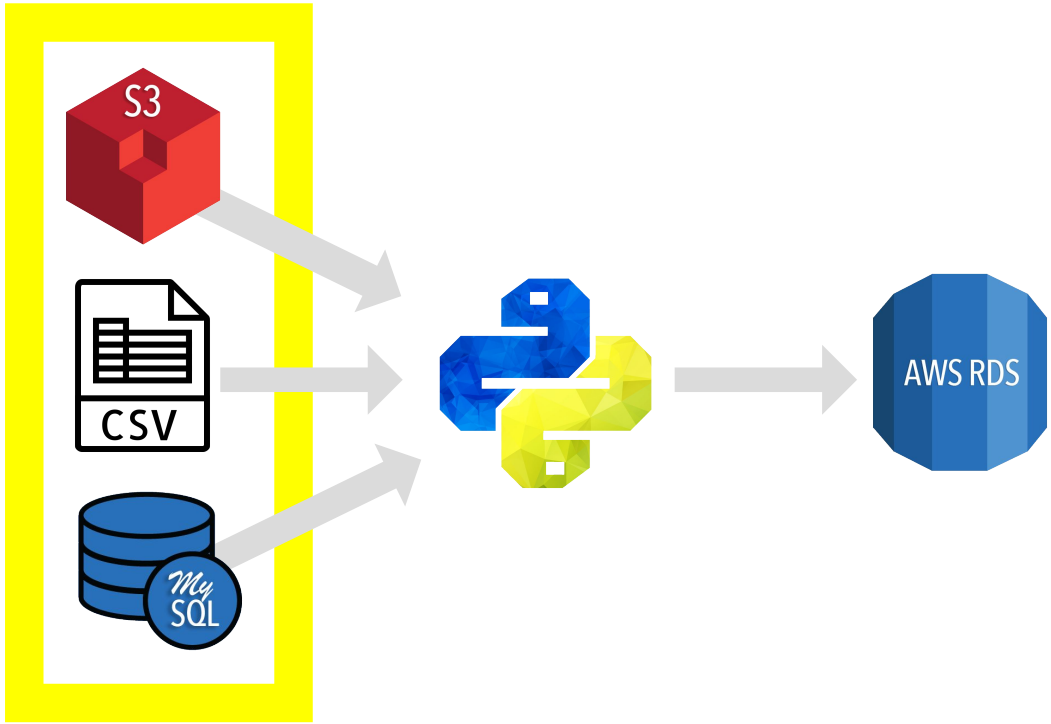
Data integration is an important part of working with data.



Extract

Data may come from disparate sources, such as:

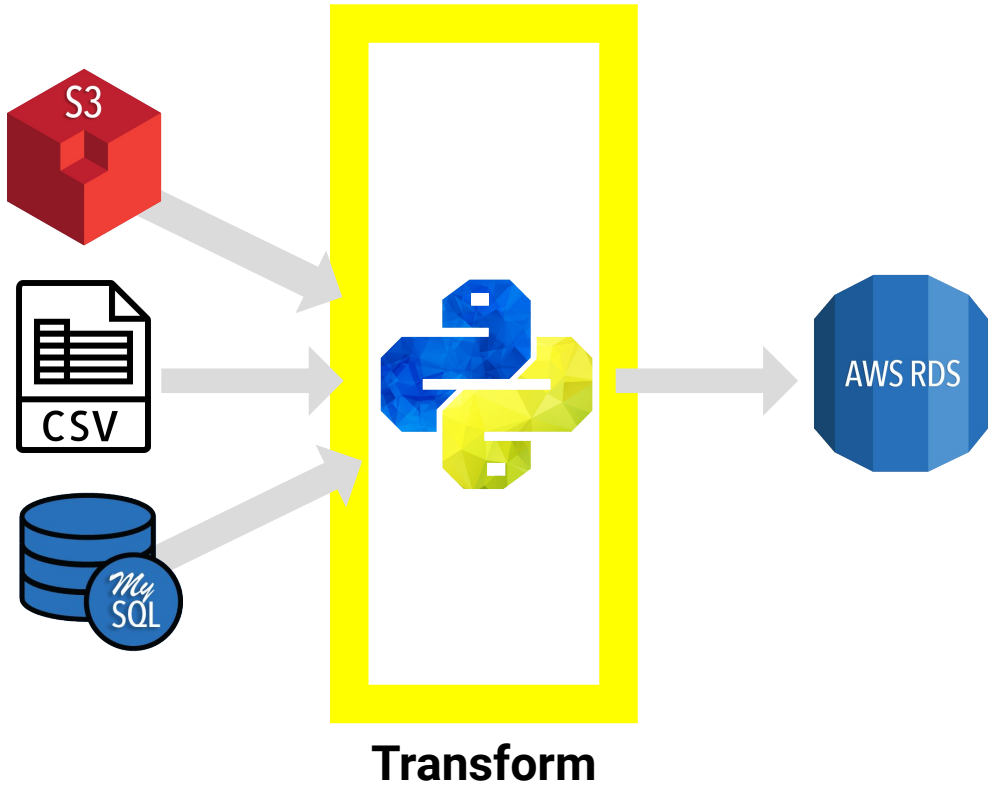
- CSV files
- JSON files
- HTML tables
- SQL databases
- Spreadsheets



Transform

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

- Data Cleaning
- Summarization
- Selection
- Joining
- Filtering
- Aggregating

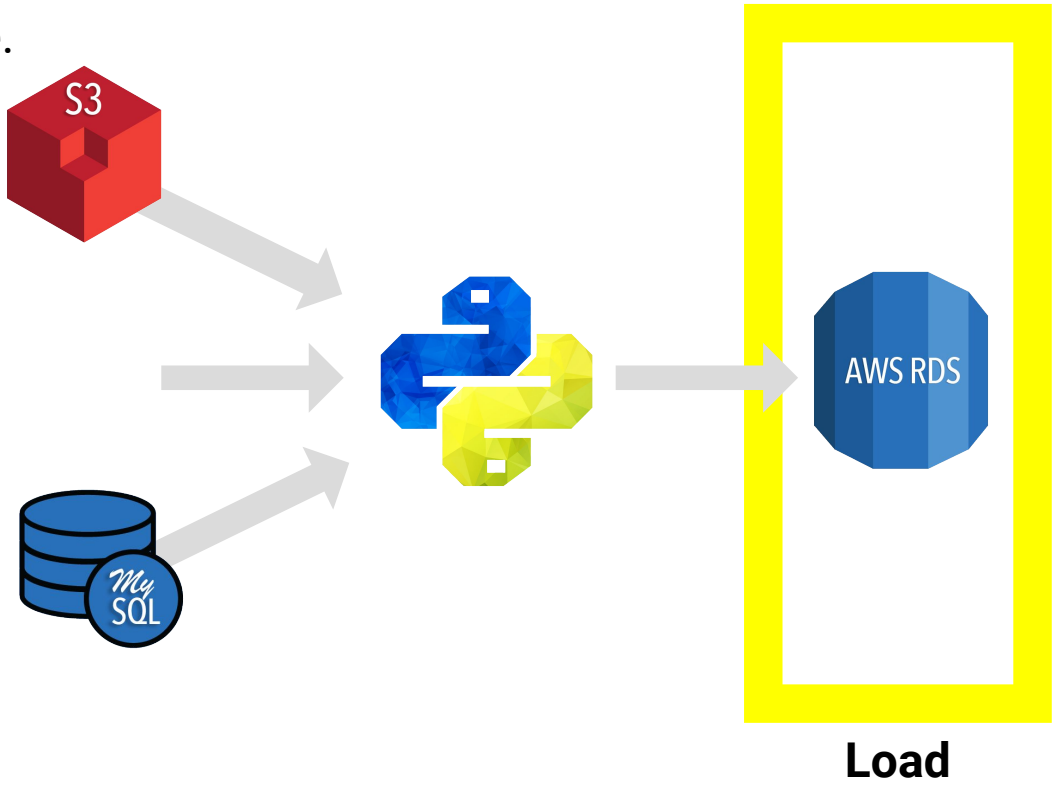


Note: We will use Python and pandas for transformation, which can also be done with SQL or a specialized ETL tool.

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 Project Requirements

Requirement	Solo/Duo	Trio
Proposal	Required	Required
Number of Sources	2	3
Source code on Github	Required	Required
Final Report	Required	Required
Flask API	Optional	Required

Project Proposal

Get this submit as soon as possible so you can get started.

- Your names
- Datasets you intend to use
- What useful investigation could be done with the final database
- Whether final database will be relational or non-relational. Why?

Types of sources

You must use at least 2 *different* types (or 3 if you're in a trio)

- Excel (or other local file)
- CSV
- PostgreSQL
- MongoDB
- Scraped Webpage
- Web API

Final Report

- Detailing the process of the extraction, transformation, and loading steps
- What data sources you chose, and why?
- Explain why you have performed the types of transformation you did
- Why you chose the type of final database
- Schema of the tables/collections in the final database
- Hypothetical use case(s) for your database

Questions to consider in your ETL

- Is my data redundant?
- Is there a way to normalize this data?
- Can I accomplish the same thing with less code?
- Is my code maintainable? If I let someone else read it, would they understand it?
- Why would someone want to use my final dataset?

Today

- Find some data, then get a proposal sent **ASAP!**
- Expect feedback. You may need to revise!
- Get started!

Questions?