ETL: Extract, Transform, Load

The Data Bootcamp

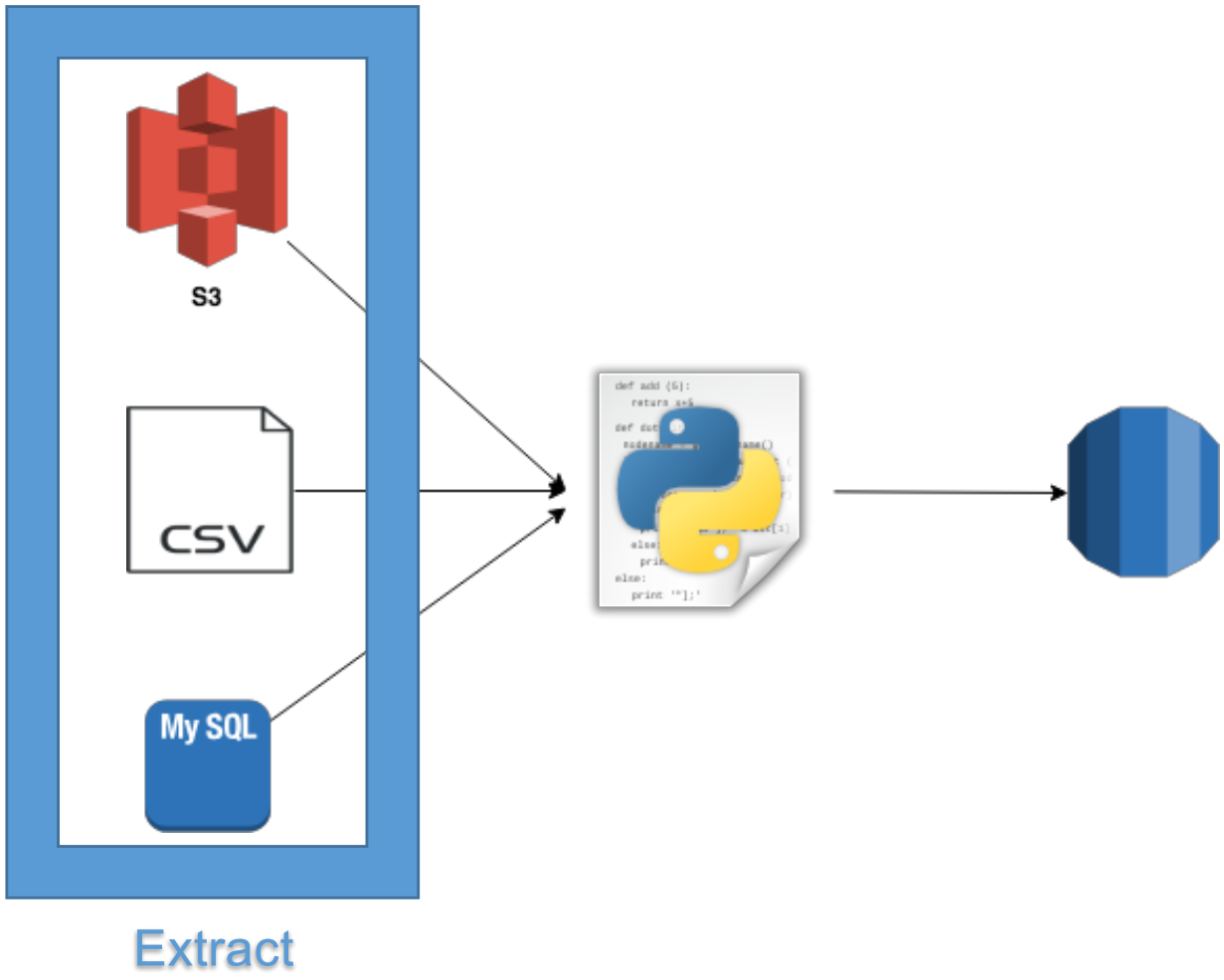
ETL

- Data integration is an important part of working with data.
- **Extract**: read the data, often from multiple sources
- **Transform**: clean and structure the data in desired form
- **Load**: write the data into a database for storage

Extract

Data may come from disparate sources, such as:

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

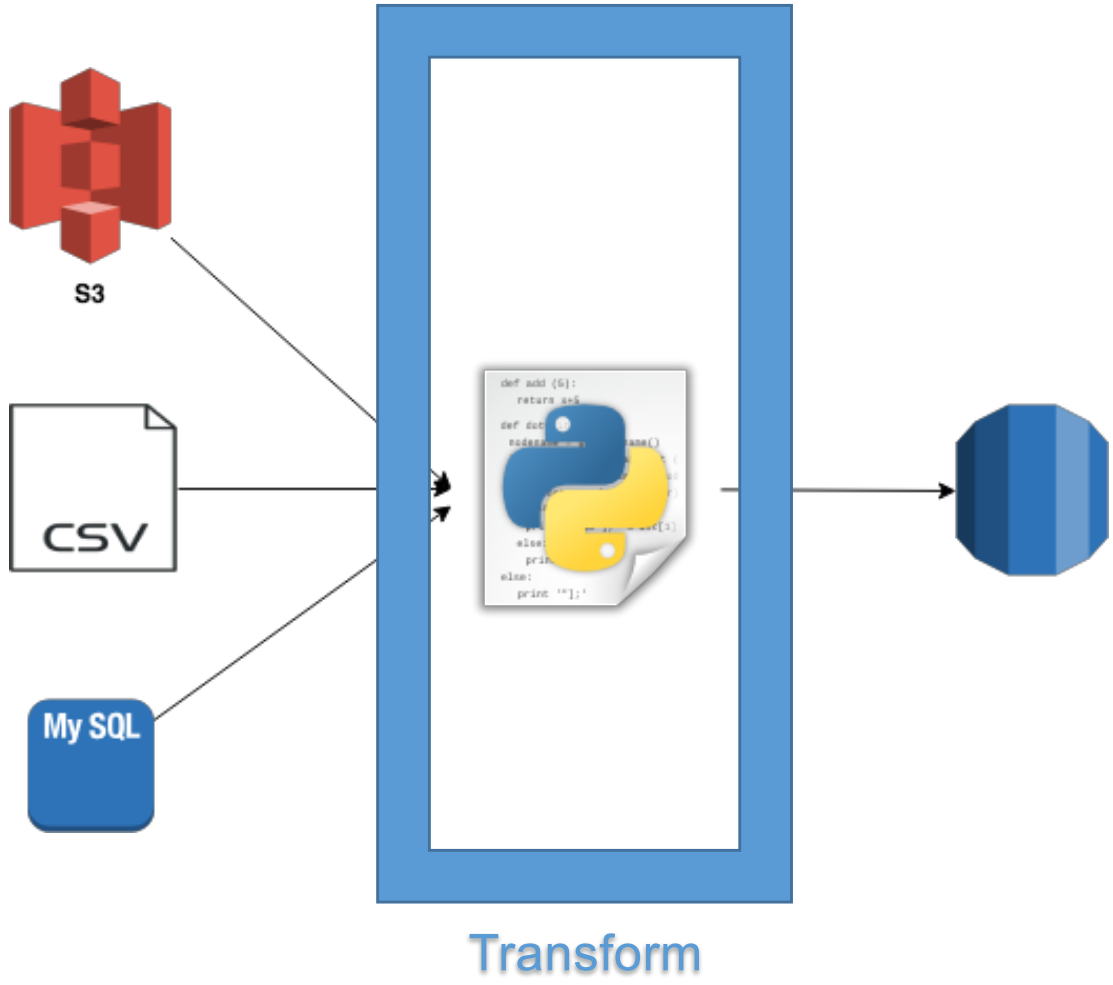


Transform

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

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

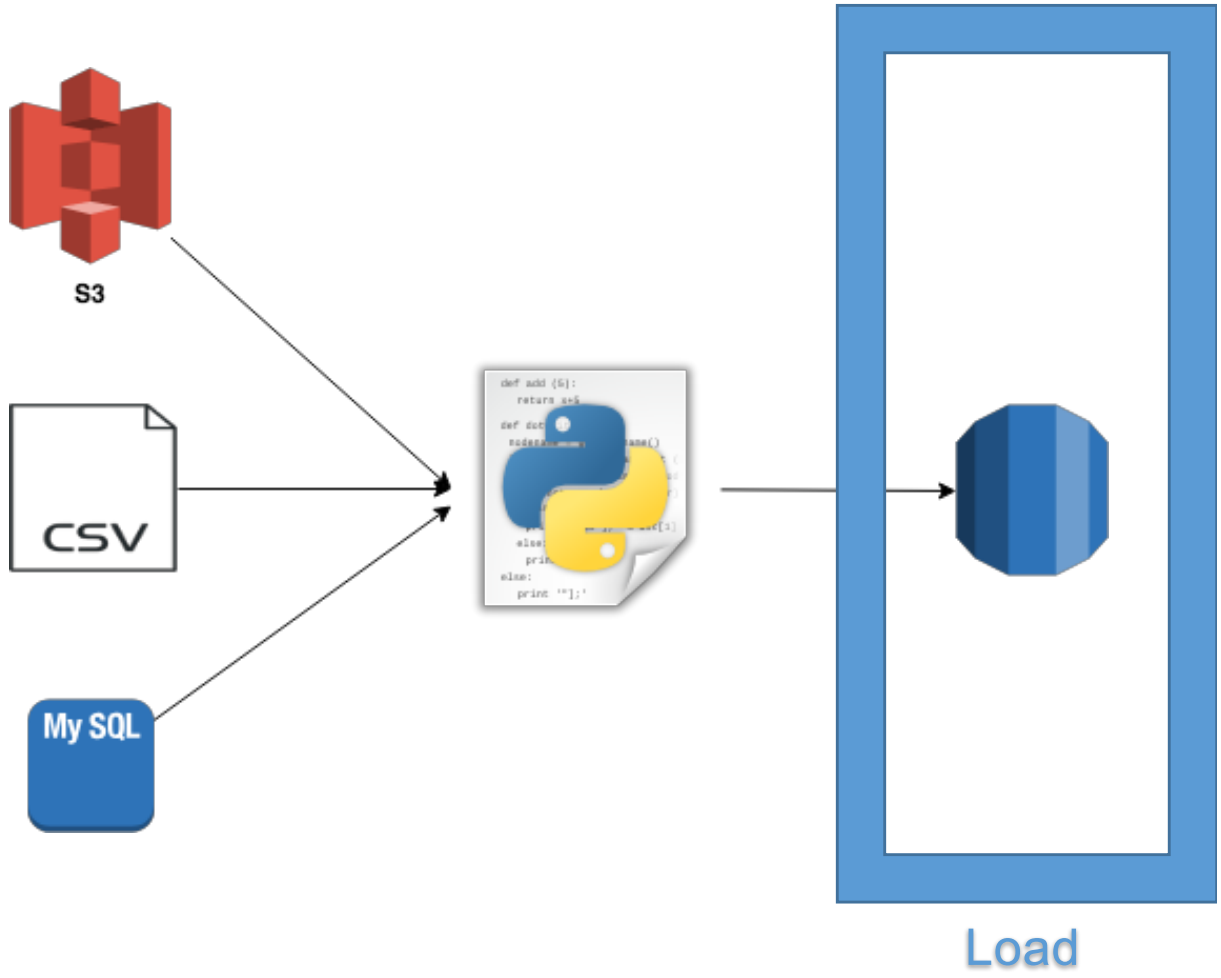
Note: We will use Python and Pandas for transformation, but this could be done with SQL or a specialized ETL tool.



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



Requirements

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

Proposal

- Your name (and partner if Duo)
- Data sets you intend to use
- What useful investigation could be done with the final database
- Whether final DB will be relational or nonrelational, and why

Types of Sources

- You must use at least two sources of *different* types

- ”Flat file”
  - CSV/TSV/DSV File
  - Excel File
- SQL Database
- Mongo Database
- Scraped Web Page
- Web API

Final Report

- Data Sources
- Detailing the process of the extraction, transformation, and loading steps
- What data sources you chose, and why
- Explication why you have performed the types of transformations you did
- Why you chose the type of the final database
- Schema of the tables/collections in the final database
- Hypothetical use cases for your database

If you finish early

- Build a Flask API (required if you are a duo group)
  - Build a Flask web API that responds with JSON of queried data in your database
  - Ask instructional staff for help if you would like to load it to a live server (we use Heroku)
- **THEN** you may work on this week's homework
- **THEN** you can start on the Julia lessons if you have time leftover! (Cam is still working on these)

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 without me being there?
- Why would someone want to use my final dataset?

TODAY:

- Find some data, then get a proposal to Cam **ASAP**
  - I like Google Docs, but will suffer your Word files.
- Expect feedback, you may need to revise!
- If you need to pivot, we can discuss that.

Questions / Discussion