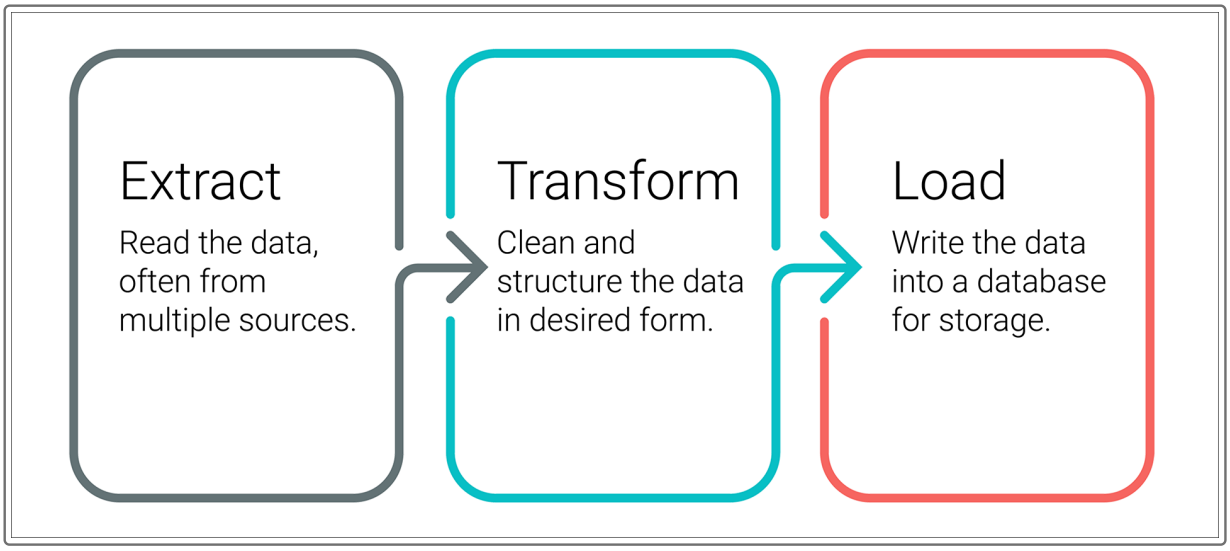
8.1.1 Extract, Transform, Load

Britta is excited to prepare for the hackathon. In data analysis, a hackathon is an event where teams of analysts collaborate to work intensively on a project, using data to solve a problem. Hackathons generally last several days, and teams work around the clock on their projects.

Britta needs to gather data from both Wikipedia and Kaggle, combine them, and save them into a SQL database so that the hackathon participants have a nice, clean dataset to use. To do this, she will follow the ETL process: **extract** the Wikipedia and Kaggle data from their respective files, **transform** the datasets by cleaning them up and joining them together, and **load** the cleaned dataset into a SQL database.

The idea behind ETL is straightforward. Raw data exists in multiple places and needs to be cleaned and structured before it can be analyzed. ETL breaks this problem into three steps, or phases: Extract, Transform, and Load.

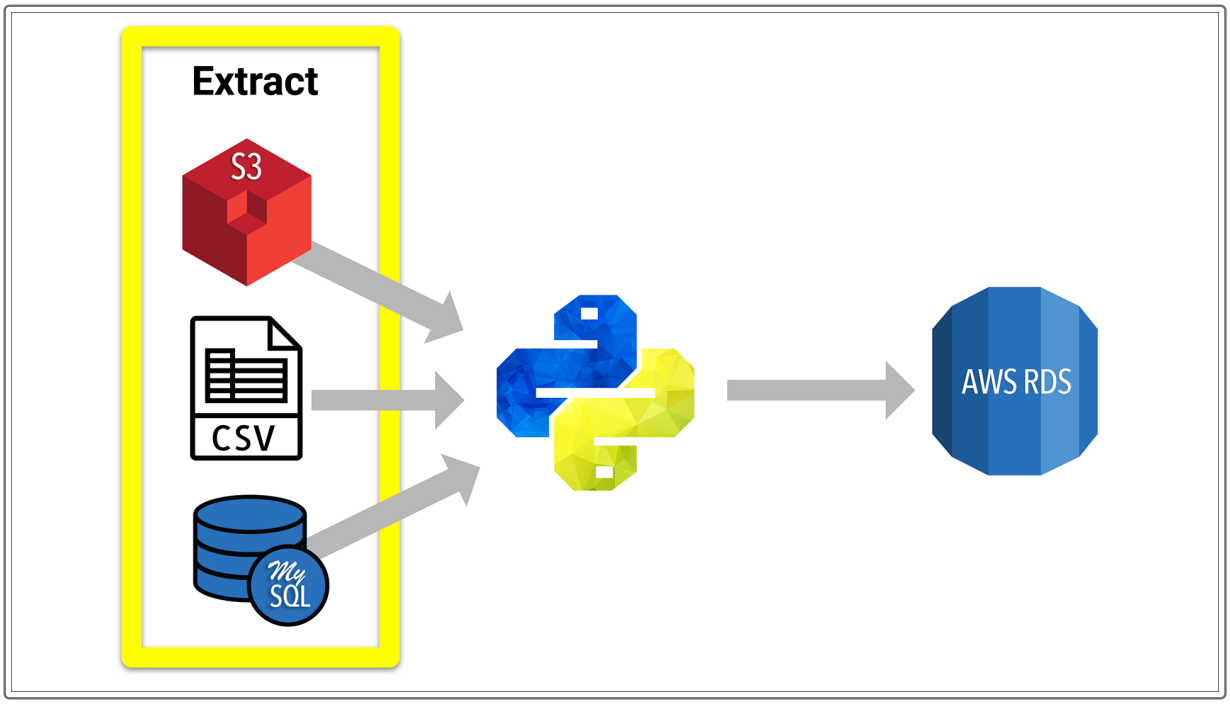


Watch the following video to get acquainted with ETL.

ETL is a flexible process for moving data. It can be as simple as a one-time migration from one database to another, or as complex as an ongoing automated collection of messy, real-time data from many different sources.

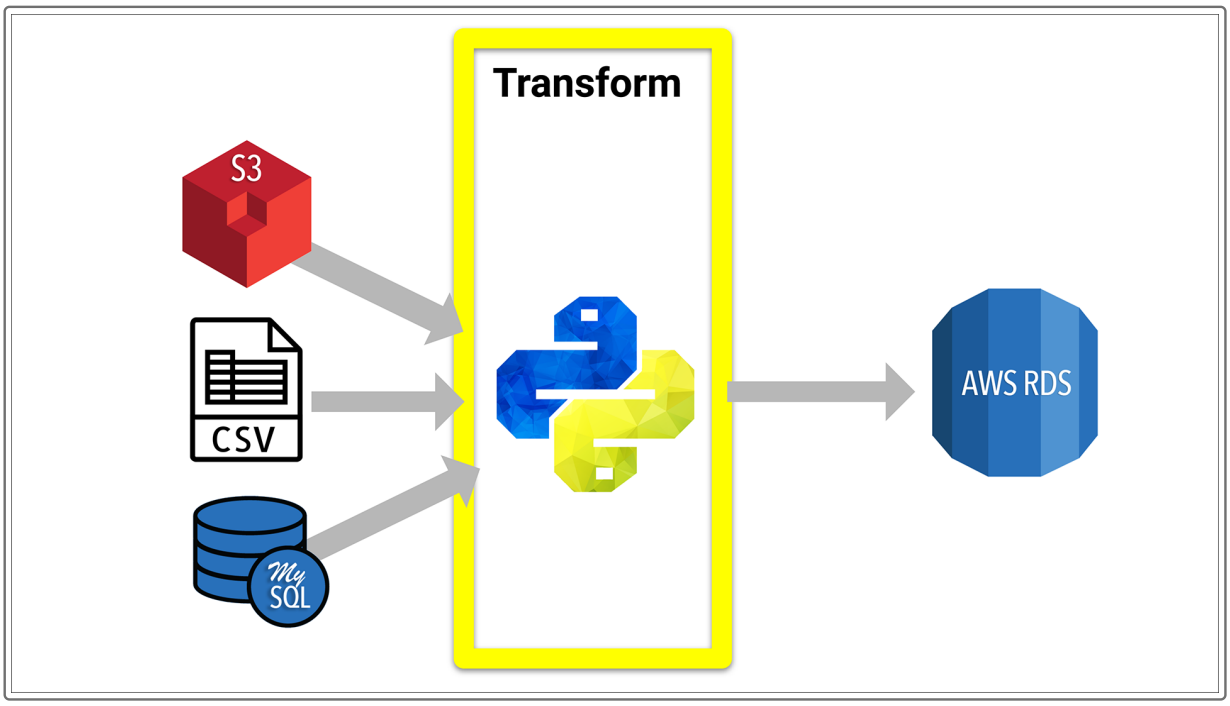
In the Extract phase, data is pulled from external or internal sources, possibly disparate. The sources could be flat files, scraped webpages in HTML or JavaScript Object Notation (JSON) format, SQL tables, or even streams of sensor data. The extracted data is held in a staging area in between the data sources and data targets.

For Britta, you'll extract scraped Wikipedia data stored as a JSON, and Kaggle data stored in CSVs.



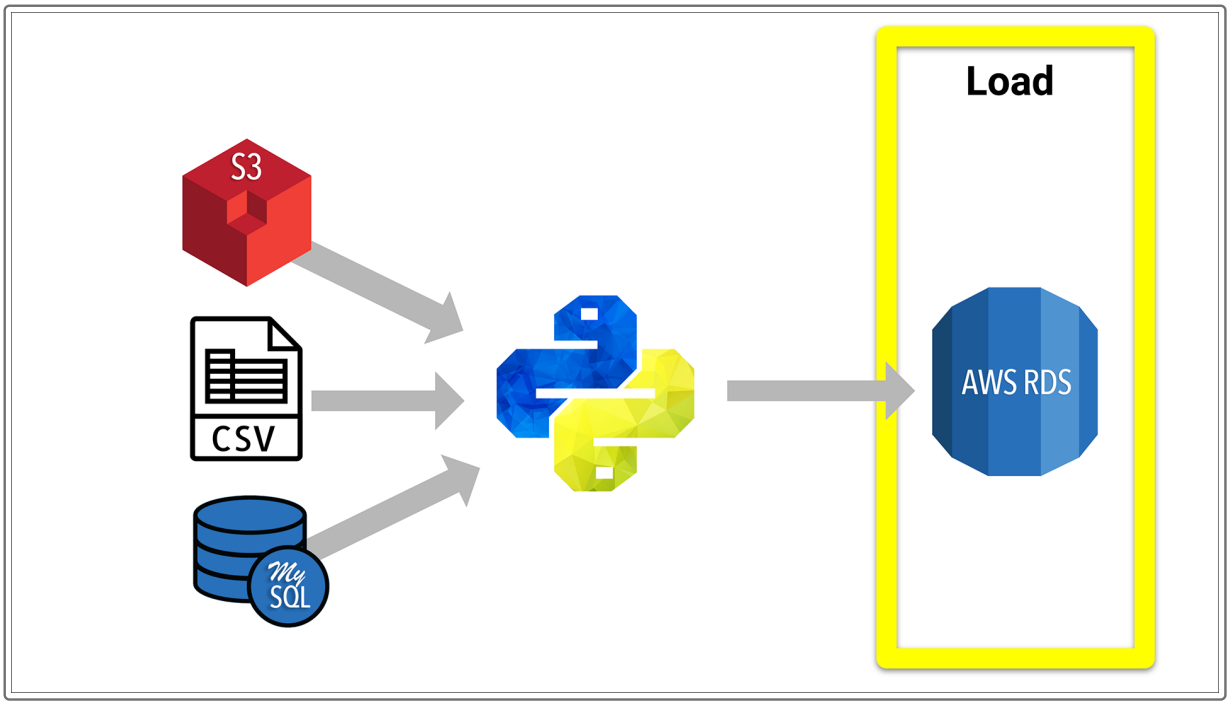
After data is extracted, there are many transformations it may need to go through. The data may need to be filtered, parsed, translated, sorted, interpolated, pivoted, summarized, aggregated, merged, or more. The goal is to create a consistent structure in the data. Without a consistent structure in our data, it's almost impossible to perform any meaningful analysis.

The transformation phase can be accomplished with Python and Pandas, pure SQL, or specialized ETL tools like Apache Airflow or Microsoft SQL Server Integrated Services (SSIS). Python and Pandas are especially good for prototyping an ETL transformation because they provide flexibility and interactivity (especially in a Jupyter Notebook), without enforcing any complicated frameworks. We will use Python and Pandas to explore, document, and perform our data transformation.



Finally, after the data is transformed into a consistent structure, it's loaded into the data target. The data target can be a relational database like PostgreSQL, a non-relational database like MongoDB that stores individual documents, or a data warehouse like Amazon Redshift that optimizes performance specifically for analytics. (We'll look at non-relational databases in more detail later.)

Britta has determined that a SQL database is the best solution for sharing the data in the hackathon, so we'll be loading our data into a PostgreSQL table. SQL databases are often the targets of ETL processes, and because SQL is so ubiquitous, even databases that don't use SQL often have SQL-like interfaces.



Now that you know the phases of the ETL process, let's get started with the first step and extract some data.