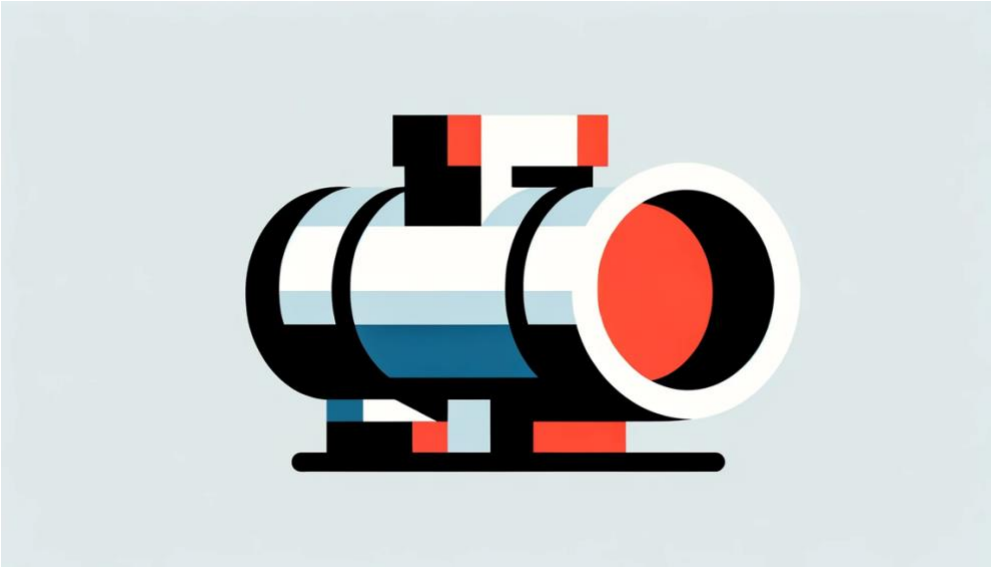
PEP 1: ETL Job



Your team has been hired by a local business that's looking for a way to import data (from periodically generated files) into a database for later analysis. You will be given a set of these source files, and your goal is to design a series of checkpoints that can be run in succession in an Extract-Transform-Load pipeline. You will have until the end of the week (Thursday afternoon) to design, develop, and prepare a 15-minute presentation to showcase the pipeline. Presentations will occur after lunch on Thursday.

Functionality:

- Checkpoint 1: Convert source files to Pandas dataframes
- Checkpoint 2: Clean source files (deal with nulls, duplicates, outliers)
- Checkpoint 3: Merge clean source files
- Checkpoint 4: Design a schema and set up the database
- Checkpoint 5: Export the data from the source files onto the database

Requirements:

- Product Requirements Document (due end of Day 1)
- Python Scripts for ETL Processes
- Database Schema (ER Diagram)

- Presentation Slides (due end of lunch Thursday)

Extensions:

- Checkpoint 6 (Bonus): Create an analysis report with findings and visualizations that describes and explores the data in the database.

Considerations when Developing a PRD

A product requirements document (PRD) is an artifact used in the product development process to communicate what capabilities must be included in a product release to the development and testing teams.

The PRD will contain everything that must be included in a release to be considered complete, serving as a guide for subsequent documents in the release process. While PRDs may hint at a potential implementation to illustrate a use case, they may not dictate a specific implementation. The process diagram below showcases the steps considered when developing a PRD.